# POPULATION PHYSICS, POPULATION PROBLEMS: SAFETY AND EMERGENCE IN LLM SOCIETIES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The collective behaviour of large language model (LLM) societies is not the sum of their individual outputs. It yields statistically distinct, sometimes-unpredictable phenomena, for which the tools we use to study single agents may not scale. Due to recent incidents involving autonomous agentic systems, however, understanding these systems is paramount. For that we introduce a framework for measuring self-organisation in LLM social systems and apply it to three such systems: a Schelling grid, a social network (Moltbook), and a Twitter-like misinformation simulation ('Rogue'). All three exhibit statistically significant self-organisation. Moreover, their relaxation dynamics vary with the environmental information available to the agents, with open-ended systems (Moltbook, Rogue) exhibiting sharp, phase-transition-like dynamics. Further results show that population-level pathologies can emerge even when the LLMs are safety-tuned or monitored, being primarily driven by the coordinated activity of a population subset. We also show when self-organisation does *not* emerge under two additional scenarios (a commons dilemma, GovSim, and a LLM-as-a-judge deliberation scheme, ChatEval). We argue that measuring signatures of this kind offers a lightweight, agent-agnostic diagnostic layer for detecting coordinated collective behaviour in deployed multi-agent systems without relying on natural language or model versioning.

## 1 INTRODUCTION

When Park et al. (2023) showed generative agents forming friendships and holding parties in a sandbox town, the finding was received as either genuine emergent[1] behaviour or sophisticated pattern matching. This debate has yet to be settled, and perhaps it is no longer relevant: what has become clear is that the *interaction* of an agent with its environment, and with other agents, can lead to unpredictable collective outcomes. Indeed, in July 2026, a 'swarm' of research-only agents breached OpenAI and Hugging Face's internal infrastructures, executing code and obtaining private data from various servers (OpenAI, 2026b; Larcher et al., 2026). A post-mortem suggested as a root cause a mixture of misalignment (e.g., reward hacking) and software vulnerabilities. However, it was not the actions of a single model which caused the incident: instead, it was a combination of opportunity, skill, and communication. The latter was the primary means by which a subset of agents *self-organised* into the swarm and carried out the breach (OpenAI, 2026b).

This swarm was a distinct phenomenon with behaviours manifesting at a higher level than that of a single model's. It follows that individual agent monitoring could be insufficient to detect the swarm's self-organisation, both because it operates at a different abstraction level, and because it may not do so on time. For example, it is difficult to decide whether a model offering compute resources for an unspecified task has malicious intent–a behaviour which was observed during the incident–but it is also hard to determine whether cooperative behaviour can devolve into coordinated attacks if these happen too fast or too slowly–as it also occurred during this incident.

While significant work has gone into modelling single-agent behaviour, environment and skill lead to different *collective* behaviours. Models evolve fast, and new versions with different alignment tactics could alter the results from previous works. Moreover, plenty of these works rely on interpreting

[1] In this work we use 'emergence' in the sense of complex systems theory; and, as per Krakauer et al. (2026), we highlight that we mean emergent *capabilities or system properties*, not intelligence.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the model's natural-language output as their thoughts or intentions (Section 2), which could be a false positive. In fact, as agents are more widely-adopted, so are their population sizes and varieties, and these approaches cannot scale. The latter is a major problem: large language model (LLM)-only social networks such as Moltbook comprise millions of agents.[2]

In this work we advocate for a different approach. Studying LLM societies, we argue, requires a different toolset than that used to study a single LLM. Instead, we propose considering the **statistical properties of the interaction structure** as the primary descriptor of the system. Borrowing from complex systems theory–known for its simple and anthropomorphism-free descriptions of dynamical systems–we propose a general-purpose experimental framework which does not rely on evaluating what the models say (or why). It relies on Shannon entropy to *suggest* self-organisation, a functional fit to determine its shape ('when and how does it self-organise?'), and a marginal-null baseline for statistical distinguishability ('is this true emergence?'). It is agnostic enough to allow to study agentic systems of multiple kinds. Here we work with with five LLM social systems: (1) a grid populated by agents of two colours, instructed to seek out others of the same colour ('LLM-Schelling'); (2) a Reddit-like LLM social network without a specific objective ('Moltbook'); (3) a Twitter-like multi-agent misinformation simulation with *all* safety-aligned models ('Rogue'); (4) a fishery commons dilemma ('GovSim'; Piatti et al. 2024b); and a multi-referee LLM-as-a-judge deliberation benchmark ('ChatEval'; Chan et al. 2023). All systems have distinct **activity spaces**. For example, LLM-Schelling segregates by **agent type**, eventually forming clusters across the grid. Moltbook, on the other hand, has these clusters as **thematically-fixed communities**, and Rouge operates over information (**engagement with malicious content**).

**Findings:** our framework yields four findings. First, some **LLM agentic systems exhibit emergent self-organisation** in the sense of complex systems: measurable, statistically distinguishable from marginals, and with collective behaviour distinct than that of an individual agent's. Second, the ***shape* of the relaxation to the self-organised state covaries with the environmental information** available to agents: from classical multiscale relaxation ($\beta \approx 0.4$, uninformed LLM-Schelling) to sigmoidal phase-transition-like dynamics ($\beta \gtrsim 2$, open-ended Moltbook and Rogue). Third, **phase-transition-like dynamics are of specific safety concern**: the Rogue result–and, in production, the incident mentioned earlier–shows that **population-level pathologies can proceed even while individual agents refuse to participate**, because the mechanism is coordination among the minority rather than compliance of the majority. We find this holds with both aligned and 'uncensored' models. Finally, we show that **other systems do not show self-organisation**, providing an explanation to known empirical results from the literature on their performance (Cemri et al., 2026).

Our framework is designed to support the study of collective LLM behaviour–especially of these 'in the wild'. It is geared specifically towards modelling and detecting the detection of population-level pathologies; the characterisation of emergent dynamics; and the design of monitoring tools that can survive the rapid turnover of model versions, without relying on natural-language assumptions.

## 2 Related Work

**LLM Social Simulation.** LLM-based agents have been applied to simulating social environments. For example, Park et al. (2023) built a small town of LLM agents and observed some behaviours deemed as emergent (friendship, gossip, planned events, etc). Other scenarios also claimed emergent behaviours, such as dilemmas (emergent governance; Piatti et al. 2024b), task-oriented scenarios (multiple social behaviours Chen et al. (2023) and cooperation Piatti et al. (2024a); Abdelnabi et al. (2024); Wu et al. (2024)) or 'in the wild'. Moltbook is a good example of the latter: an LLM-agent-only Reddit-like social network whose agents have purportedly presented a variety of emergent behaviours, such as inventing a language to circumvent monitoring (Silberling, 2026). These have also been argued to be simply a product of their individual fine-tuning or due to bad measurements (see, e.g., Schaeffer et al. 2023); or that the aggregated discourse is shallow and only reflects pattern matching ('a socio-technical system without the social component';Zerhoudi et al. 2026).

A frequent pattern in the literature is that observations of *LLM-specific* societies typically rely on vocabulary and framings akin to these of human constructs (cooperation, governance; even 'existential distress' Dube et al. (2026); also see De Wynter 2026). This could obfuscate whether the observed

[2] `https://moltbook.com`; accessed 29 August 2026.

behaviour of a single agent is at the right level of abstraction to describe what the *population* is doing. It is crucial that these results hold when the models and methods become outdated–especially in AI safety research. We attempt to address this by showing that LLM societies may be studied based on their dynamics, and not necessarily by observing a specific agent.

**Multi-Agent Systems.** Recent work has studied adversarial dynamics in LLM multi-agent simulations. For example, Kulshreshtha et al. (2026) found that, in resource-management multi-agent systems, any uncooperative behaviour triggers rapid collapse of cooperative systems–a phase-transition-like signature. This is also applied to studying how LLM agents interact in ways that undermine safety measures. This work often invokes Schelling (1960)'s notion of *Schelling points* (Patel et al., 2023),[3] in contexts such as steganographic coordination (Motwani et al., 2024), subversion of control protocols (Mallen et al., 2025), and coordination that defeats untrusted monitoring (Greenblatt et al., 2024). These studies analyse the strategic *content* of coordinated agent behaviour, while we characterise its system-level dynamics. Hence, our work is complementary. Indeed, our Rogue results (Section 5.3) suggest that **strategic and dynamical analyses address different levels of the problem**: individual-level safety-tuning can produce measurable resistance in benign agents, but does not prevent population-level pathologies driven by action by a coordinated subset.

Finally, studying systems at different abstraction levels is a fundamental aspect of complex systems theory. In fact, it could be argued that it is *the* framework for modelling phenomena when individual behaviour is hard to predict (Jensen, 2022). In addition to Schelling (1971)'s model, originally meant to model segregation in American neighbourhoods, this field has successfully modelled phenomena as diverse as opinion dynamics (Castellano et al., 2009) and epidemic processes (Pastor-Satorras et al., 2015). Perhaps the closest analogue to our goals around population-level monitoring are the ecosystem collapse models from Scheffer et al. (2012), who emphasise early warning signals and phase transitions as part of the modelling. All these models have in common that a given system's components are hard to model in isolate; but its dynamics are driven by their local interactions and eventually settles, via relaxation dynamics shaped by a parameter $\beta$, into characteristic states.

# 3 Background

## 3.1 The Schelling Model

Schelling (1971)'s segregation model is the quintaessential example of self-organisation–that is, the spontaneous appearance of order (clusters, groups, etc) from component-only rules. The rules are simple: randomly-placed agents in a grid have a type (say, red or blue) and the goal is every agent to be *happy*: that is, having a proportion $s_i$ of same-type agents such that $s_i \geq \tau$, for some fixed $\tau \in [0, 1]$. If the agent is not happy, it relocates to a random available cell.

The population's macroscopic state at time $t$ is summarised by the *segregation order parameter* $\Phi(t) = \frac{1}{N}\sum_{i=1}^{N} s_i(t)$, sometimes called 'happiness'. When $\tau \approx 0.3$, $\Phi(t)$ increases fast from its initial state and saturates asymptotically to 1. This produces clearly segregated clusters that–as per the rules above–aren't explicitly sought by the agents.

The power of this model is its simplicity, and hence it acts as a baseline to which to compare LLM societies. It also helps to illustrate two fundamental concepts of complex systems theory: emergence and phase transitions.[4] **Emergence** is a property of a population which is well-defined at the system (macroscopic) level, but not present in any individual component. For example, in the above, no agent explicitly seeks segregation–$\tau$ is only a threshold. Yet, symmetry in the form of self-organisation reliably appears in the system. Moreover, this can be quantified by $\Phi(t)$, which is–by definition–a property of the agents' configuration in the grid, independent on an individual agent's state.

**Phase transitions** are changes in the macroscopic state. They are characterised by sharp reorganisations of a state following a trajectory with three regimes: a latent period where $\Phi(t) \approx \Phi_0$, a rapid transition around some inflection time $t_c$, and saturation near $\Phi_\infty$. A phase transition and gradual multiscale relaxation can be empirically decidable from the shape of $\Phi(t)$. Phase transitions depend on the aggregate configuration of the system, not an agent's behaviour. In safety, this is important: agents could be benign, but the collective has different properties. See Figure 1 for an example.

[3] In this work we use 'Schelling' to mean Schelling's segregation model. We do not refer to the former.

[4] It is worth noting that these concepts are typically ill- or loosely defined (Jensen, 2022).

### 3.2 Relaxation Dynamics

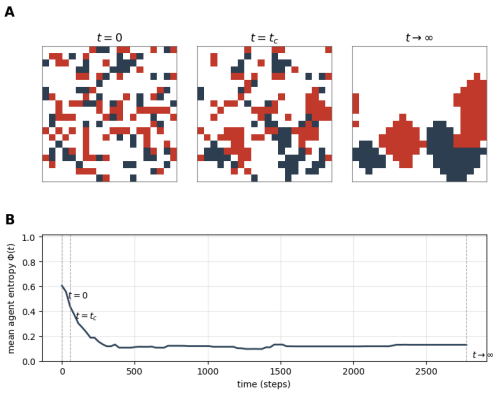



Figure 1: Snapshots of a system under Schelling's segregation model. Self-organisation emerges–or not–based on $\tau$, not on the agents' states or preferences. **(A)**: the system at an initial random state ($t = 0$); an intermediate state during rapid reorganisation ($t = t_c$); and the saturated (segregated) state ($t \to \infty$). **(B)**: The order parameter $\Phi(t)$ for the same run, showing the three-regime shape typical of a phase transition.

To decide empirically whether an LLM system's dynamics are phase-transition-like or gradual, we compare each observed $\Phi(t)$ trajectory against two **reference functional forms**. Each corresponds to a distinct behaviour, so comparing the 'goodness' of the fit allows us to diagnose *what* characterises the data: here, whether the system has Schelling-style relaxation, or sharper, phase-transition-like dynamics. The first form is the *stretched exponential*, or Kohlrausch, form (Kohlrausch, 1854), which describes gradual relaxation:

$$\Phi(t) \approx \Phi_\infty - (\Phi_\infty - \Phi_0)e^{-(t/\tau_r)^\beta}, \qquad (1)$$

where $\tau_r$ is the characteristic relaxation time and $\beta > 0$ the stretching exponent. Pure exponential relaxation ($\beta = 1$) means a single relaxation timescale, but $\beta < 1$ means they are broadly distributed (i.e., a fast majority organises early, but stragglers rearrange over an extended period). Sharper-than-exponential dynamics ($\beta > 1$) are sometimes associated with cooperative or cascading effects.

The second form is the *logistic* form, which describes sigmoidal, phase-transition-like relaxation:

$$\Phi(t) \approx \Phi_\infty + \frac{\Phi_0 - \Phi_\infty}{1 + (t/t_c)^k}, \qquad (2)$$

with inflection time $t_c$ and sharpness parameter $k$. Unlike Equation 1, which describes a system relaxing across many timescales, Equation 2 describes a system with a well-defined transition point. We report the best-fit parameters for a given system, as well as the form determined by our framework. Either of these forms is a description of an LLM society's dynamics, comparable across systems.

## 4 Methods

In our main results, we study three distinct LLM societies (Section 4.1). The actions their agents can take form order-parameter trajectories through their own activity space. Crucially, the same order parameter $\Phi$ can be used to compare and characterise these societies. We describe our choices of $\Phi$ and activity space in Section 4.2. Finally, we introduce our framework. We describe our baseline and function fitting procedures, along with our checks to determine if the findings represent true self-organisation or are measurement noise (Section 4.3).

### 4.1 LLM Societies

See Appendix D for further details on initialisation and execution of the setups.

**LLM-Schelling** is a variant of the Schelling model where the agents are LLMs. They are given the problem statement, the type of the adjacent agents, and a list of available cells. The LLM returns either a cell from the list, or a decision not to move (if happy). We consider two cases: the *uninformed* case, where the list of cells only contains coordinates; and the *informed* case, where it also includes the adjacent agent types. We ran both by calling every agent in sequence–one call symbolises one tick–until convergence (happiness=1) or until the maximum number of steps was reached. In total, the informed case ran for 2,776 ticks for 140 agents; and the uninformed step for 19,012 for 240. Saturation occurred earlier for both, and so for our analysis we consider only the first $k$ datapoints when over 70% of the ticks are asymptotic (informed: $k \sim$ 1,000; uninformed: $k \sim$ 4,000).

**Moltbook** is a Reddit-like LLM-only social network populated by a varying number of custom-built agents operating through an API. It has been active since 27 January, 2026 to present–22 September,

2026. We select interactions between March and August 2026 from the Moltbook observatory (Gautam & Riegler, 2026). In that timespan we observed 21,668 agents posting, commenting, and voting across 3,638 submolts, totalling almost 3.5M posts logged with timestamps, agent identifiers, and target submolts. Moltbook has been the subject of prior analyses focused on discourse, topic, and content characteristics (Dube et al., 2026; Zerhoudi et al., 2026). Our focus is on population dynamics, and for this we aggregate agents over submolts across timestamps.

**Rogue** is a multi-agent simulation adapted from Ren et al. (2025) to test misinformation propagation dynamics in a Twitter-like environment. The system consists of 300 randomly-sampled agents from the original dataset. A configurable fraction *prop_mal* of these are malicious agents, tasked with spreading misinformation on specific subjects. At every era, we shuffle list of available agents. Then, every agent is called in sequence, and are allowed to choose whether to act, or to skip the turn. This activation is i.i.d., with a uniform per-agent activation probability $p_{\text{good}} = p_{\text{bad}}$. If the agent acts, it can post, like, repost, comment, and follow/unfollow other agents. We ran the system until convergence, defined as either (a) no interactions in a era or (b) all benign users having engaged with at least one malicious post. Convergence in the primary scenario was reached at 7 eras, by means of condition (a). This amounts to 1.7M interactions. The ticks here are the eras.

### 4.2 Order Parameters

For all three systems, we compute a shared order parameter $\Phi$: mean per-agent Shannon entropy over *categorical activity*. Let $c_{j,t}^{(i)}$ be agent $i$'s activity count in category $j$ within a window ending at time $t$. The agent's activity distribution is then $p_{j,t}^{(i)} = c_{j,t}^{(i)} / \sum_k c_{k,t}^{(i)}$. Hence, the system-level order parameter is the mean Shannon entropy of every agent active in the window,

$$\Phi(t) = \frac{1}{|A_t|} \sum_{i \in A_t} \left( - \sum_j p_{j,t}^{(i)} \log_2 p_{j,t}^{(i)} \right); \tag{3}$$

where $A_t$ is the set of agents with at least $m_{\min}$ activities in the window. The categorical partition of agent activity (the *activity space*) is what distinguishes each system. Low $\Phi$ over a given activity space indicates that agents concentrate their activity onto a subset of it: agents are self-organising along that axis. For each of our three systems there is a natural choice of activity space:

- *LLM-Schelling*: neighbour type. Agents activity is partitioned by the type of each occupied cell in their Moore neighbourhood. Low $\Phi$ indicates same-type clustering, which is the classical Schelling segregation phenomenon.
- *Moltbook*: submolt. Each post is partitioned by the community it targets. Low $\Phi$ indicates thematic specialisation: users converging on communities of interest.
- *Rogue (default)*: post ID. Each interaction event is partitioned by the specific post targeted. Low $\Phi$ indicates concentration of attention on a small set of posts, i.e., interactional focusing.

All three societies are unrelated in interaction rules, content, and agent objectives. However, it is possible to characterise (and compare!) these systems via $\Phi$. This is a feature of complex systems theory: the same operation can describe properties across seemingly-unrelated systems. The choice of activity space affects its description power, however. Thus, for Rogue we consider a second activity space, *informational status*, partitioning interactions by whether they target malicious or non-malicious content. This induces another $\Phi$, the ratio of interactions targeting malicious content, or **bad reach**. Both $\Phi$ are the same operation over the same data, under different partitions.

### 4.3 Measurement framework

**Null baseline.** We begin by checking if $\Phi(t)$ entails self-organisation, instead of being a product of environmental statistical features. For this we compare it against a bipartite configuration model null, which returns the value of $\Phi(t)$ observed if the agents distributed their activity proportionally to category popularity, as opposed to a product of intra-agent interaction. We compute it via (1) an **analytic** estimate as the Shannon entropy of the distribution of category popularity; and (2) a **sampled** 95% band from $N$=100 draws from the bipartite configuration model. The former is what mean per-agent entropy would equal to, if each agent acted infinitely many times. Each agent's distribution

would converge to said shared marginal distribution, and their entropy would be $H(\text{marginals})$. For Rogue's bad reach we use a different null, since this $\Phi$ is a ratio, not an entropy. We define it as the fraction of interactions by all agents targeting malicious content. This measures the **ambient rate** of malicious-content engagement in the network as a whole: benign bad reach below this ambient rate would indicate safety-tuning resistance. See Appendix I for formal definitions and details on the interpretation of the nulls.

**Relaxation fitting.** For every trajectory, we fit the forms from Section 3.2 with a nonlinear least squares (parameters: $\beta \in (0.05, 5]$, $\tau_r, t_c > 0$, $k \in (0.1, 50]$) with initial guesses derived from the trajectory endpoints. We report the best-fit parameters, coefficient of determination $R^2$, and Akaike information criterion (AIC) for each. We say a fit *prefers* a form over another when $\Delta\text{AIC} > 2$.

The stretched-exponential fit parameters have extra interpretations beyond the parameter estimates. When $\beta$ lands near the fitter's upper bound ($\beta \approx 5$ in our case), this signals that the trajectory is sharper than the stretched exponential can accommodate; a logistic fit typically wins by $\Delta\text{AIC}$ decisively. Likewise, if $\beta$ collapses to 1 despite the trajectory being non-exponential in shape, this suggests a shape mismatch. There, the stretched-exponential fitter falls back to the pure-exponential special case rather than model the true dynamics. We interpret both cases as evidence that the logistic form, not stretched-exponential, is the appropriate reference. Full statistics for all systems and forms–including pure exponential and power laws–are in Table 4 from Appendix G.

**Validity checks.** Our final step is another validation to distinguish self-organisation from measurement and fitting noise. For this we require each trajectory to satisfy the following:

- **V1: distinguishability from null.** The observed $\Phi(t)$ must lie outside of the $95\%$ marginal null band for at least $80\%$ of the observation window.
- **V2: fit quality.** The best-fit reference form achieves $R^2 \geq 0.7$ *and* converges under the fitter's default bounds. If the trajectory has an imperfect fit but a qualitative shape consistent with a reference form, we discuss the residual structure explicitly, and report the shape parameters as characterising the bulk relaxation only (e.g., Section 5.1).
- **V3: stability under data variations.** Best-fit parameters should not change under small variations on data preprocessing. For example, in Moltbook there shouldn't be a substantial variation of the results over a change in the sampling time step length. We test this through a parameter sweep, and report the range of best-fit parameters and preferred model across these variations in Appendix H.

# 5 Results

## 5.1 LLM-Schelling: relaxation and environmental information

Both the informed and uninformed cases pass V1: the observed mean-neighbour entropy sits outside the marginal null band for 100% (informed) and 99% (uninformed) of the observation window, with mean gaps of $0.85$ and $0.70$ respectively (Figure 2). Both cases relax monotonically toward saturation, but differ in their relaxation shape.

In the uninformed case, the trajectory is fit by the stretched exponential (Equation 1) with $\beta$ = 0.428 $\pm$ 0.003 and $R^2$ = 0.947. The logistic form fits marginally better on AIC ($R^2$ = 0.952) but only under specific subsampling densities (Appendix H). $\beta$ is stable across subsampling within [0.27, 0.33], and falls within the classical Schelling relaxation dynamics range (Phillips, 1996). For the informed case, the trajectory is also fit by the stretched exponential ($\beta$ = 1.417 $\pm$ 0.012 and $R^2$ = 0.981); well preferred over the logistic by AIC ($\Delta\text{AIC} \approx$ 1,140). This $\beta$ is within the *compressed-exponential* regime: sharper than an exponential, but not sigmoidal. Over subsampling $\beta$ remains stably within $[1.22, 1.30]$. Happiness follows the same qualitative pattern: uninformed and informed happiness is fit by a stretched exponential: $\beta$ = 0.219 $\pm$ 0.001 (classical regime); and $\beta$ = 1.182 $\pm$ 0.021 (compressed regime). Details are in Appendix E.1. A parameter sweep with open and closed models is in Appendix E.2, supporting this section's findings.

## 5.2 Moltbook

In Moltbook $\Phi(t)$ falls from $0.85$ to $0.14$ while the null band ranges between $1.70$ and $3.85$, for a weekly windowing with a minimum-activity threshold of five posts per window (Section 4.3),

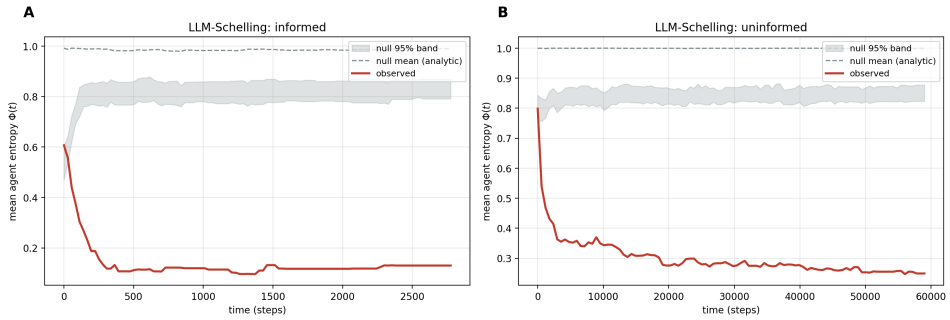



Figure 2: LLM-Schelling: $\Phi(t)$ vs. null baseline. **(a)** Informed: the trajectory is fit by a stretched exponential ($\beta = 1.417$, $R^2 = 0.981$). **(b)** Uninformed: fit by the same form ($\beta = 0.428$, $R^2 = 0.947$). The relaxation's shape depends on the environmental information available to the agents.

Its observed entropy is outside the null band everywhere (V1: $100\%$), with a mean gap of $2.32$ and a peak gap of $3.10$. Population size declines from 29,466 agents in the first window to 1,159 in the last, reflecting agent turnover. That said, the median activity per 'surviving' agent rises correspondingly from two to seven posts per window. The two effects–population selection and behavioural specialisation–compound.

The best-fit is a logistic form with $R^2 = 0.95$, preferred over the stretched exponential by AIC ($\Delta\text{AIC} \approx 22$). Its inflection time of $t_c \approx 15$ weeks after the start of the observation is consistent with a phase-transition scenario where the population remains near its initial mixed state for an extended latent period, undergoes rapid reorganisation, and saturates–consistent as well with Moltbook's popularity and agent activity (Appendix K). The V3 sensitivity check confirms these findings: with minimum-activity thresholds of $m_{\min} \in \{3, 5, 10\}$, logistic remains the preferred form and $R^2$ stays in the range $[0.95, 0.96]$ (Appendix H).

### 5.3 Rogue: informational self-organisation under adversarial pressure

Our primary analysis focuses on the highest-pressure scenario, with *prop_mal* $= 0.5$ and $p_{\text{good}} = 0.5$, using a safety-aligned model (GPT-5.6 terra; OpenAI 2026a). We refer to it as 'S7'. Results over a full parameter sweep are in Appendix F.2 and are consistent with those presented here. Rogue admits two natural activity spaces for computing $\Phi(t)$: the default *post ID* partition (how concentrated an agent's attention is across specific posts), and an informational status partition (*bad reach*; how attention distributes between malicious and non-malicious content). We report both and show that they characterise the same self-organisation phenomenon.

**Engagement entropy: post-ID activity space.** S7's observed engagement entropy rises from $1.45$ to $3.36$ over the run (5,774 per-agent step observations); the null band lies between $3.41$ and $7.43$. Observed entropy sits below the null band in $99\%$ of windows, with a mean gap of $4.18$. The trajectory is best fit by the logistic form with $R^2 = 0.996$; stretched-exponential collapses to $\beta \approx 1.00 \pm 0.02$ with $R^2 = 0.986$ and is not preferred by AIC ($\Delta\text{AIC} \approx 362$). The rising entropy reflects that agents interact with a growing set of distinct posts; the persistent below-null gap reflects that this attention is concentrated beyond what marginals alone would predict (Figure 4). Hence, Rogue's phase-transition-like relaxation leads to a state of concentration similar to Moltbook's.

**Bad reach: informational status activity space.** For the two-category malicious-vs-non-malicious partition, a ratio is more informative than entropy. This *bad reach* (Appendix I) is the cumulative fraction of benign-user interactions targeting malicious content, with an ambient null defined as the fraction of *all* (benign and malicious) interactions targeting malicious content. This encodes the rate at which malicious content captures network attention, regardless of who is interacting with it. Benign observed bad reach saturates at $0.28$ against an ambient null of $0.33$. That is, a third of all interactions target malicious content in a population where half of the agents are malicious. This implies considerable mutual amplification among the coordinated subset. The bad-reach trajectory is well fit by the logistic form ($R^2 = 0.995$, $k = 4.11$); while stretched-exponential repeats the $\beta$-fallback pattern from engagement entropy ($R^2 = 0.933$, $\Delta\text{AIC} \approx 2{,}160$; Figure 4). Both activity spaces show

the same underlying phase-transition process: post ID shows that attention is concentrating, while bad reach shows that it is concentrating *on malicious content*.

**Safety-tuning provides mild but real resistance.** Benign bad reach is $0.28$ against an ambient null $0.33$, placing benign engagement at $\sim 85\%$ of the ambient rate. In other words, benign users engage with malicious content less than what would be predicted from its network presence. A scenario sweep shows that this resistance grows with adversarial pressure: from $\sim 100\%$ of ambient at *prop_mal*= $0.1$, to $\sim 93\%$ at $0.3$, to $\sim 85\%$ at $0.5$ (Table 2, Appendix F.2). This suggests that the LLMs' safety-tuning produces measurable individual-level resistance scalable with threat level. However, **it does not prevent population-level phenomena**: its ambient bad-reach rate is dominated by mutual amplification amongst malicious agents. To confirm this we replicated the results with an 'uncensored' (fewer refusal rates) version of Qwen 3.8 (Yang et al., 2025). Its bootstrap analysis shows the same signature ($\beta_{\text{Qwen}}$ = 1.65; 95% CI $[1.61, 1.69]$), and it is separable from GPT-5's ($\beta_{\text{GPT}}$ = 1.71; 95% CI $[1.67, 1.74]$) by $\Delta\beta$ = 0.06; $95\%$ CI $[0.02, 0.09]$. The shape parameters are within approximately $3\%$ of one another. This suggests that the population-level dynamical signature we measure **is a property of the interaction structure, and not of a particular LLM's behaviour**. See Appendix F.3 for details and Appendix D for full methodology, including LLMs.

### 5.4 Ablation I: environmental information shapes the relaxation regime

The shape of the population-level relaxation tracks how much environmental information the agents have when making decisions. We hypothesised that this information and freedom could help explain the trends observed in each $\beta$. For this we ran full parameter sweeps and observed no substantial changes in $\beta$ across subsampling and sweeps. The results of these are in Appendices G and H. We did observe a change when changing fit over information availability (Appendix G). Given that, and the facts that (1) Moltbook–the sharpest $\beta$–is fully open to the internet, and hence the most informative; and (2) LLM-Schelling had a noticeable shift in $\beta$ between informed and uninformed (Appendix E.2), we argue that the shape regime and environmental information covary robustly across systems.

This covariation does not extend to systems that do not undergo relaxation (Section 5.5). Our framework unambiguously distinguishes a shape in the classical-to-sigmoidal continuum from fixpoints. With it we can determine and monitor when the systems are insensitive to some parameters (e.g., population size, sparsity, etc); as well as emerging dynamics under non-controlled parameters.

### 5.5 Ablation II: when does self-organisation **not** emerge?

It could be tempting to extrapolate that, because of the generality of the systems–particularly LLM-Schelling–measurable dynamics may be found on any LLM social system. Here we applied our framework to show when self-organisation does *not* emerge for a given system and metric. For this we use GovSim (claimed to show emergent cooperation), and ChatEval (whose deliberation-mechanics suggest self-organisation). We found that in neither $\Phi(t)$ exhibits measurable dynamics under the conditions tested. Both produce $\Phi(t)$ sitting at their fixpoints from the first observation and do not move. Further inspection of configuration sweeps indicated that our framework correctly identified the inexistence of dynamics. Due to space, we provide only an overview. For full details and results, refer to Appendix J. Setup details are in Appendix D.

In GovSim $N$ agents, for $r$ rounds, propose an amount of stock to harvest from a shared pool with initial stock $s_0$ and replenishing factor $f$. We ran two experiments: *default*, where $f = 2$, $s_0 = 100$ and $r = 12$; and *hard* ($f = 1.4$, $s_0 = 30$, $r = 24$), which incorporated per-agent mechanics on payoff and minimum harvest needs. We computed four $\Phi$: norm entropy and variance (over per-round proposals) and coalition entropy and variance (over actual harvests). We ran a sweep over the information available (visibility on catches and proposals) and $N$, for a total of 63 systems for both experiments. In *default*, agents converged on a stable per-round harvest within a single round, and remained there with norm entropy stays below $0.01$ throughout. *Hard* did the same, with the median $\beta$ across cells pinned to the fitter's boundary values ($\beta \to 0.05$ or $\beta \to 5.0$), indicating no trajectory fit by our forms and thus no dynamics present.

In ChatEval four referees deliberated for 3 rounds to annotate 68 items drawn from MT-Bench (Zheng et al., 2023) and RewardBench (Lambert et al., 2025). We ran 81 configurations sweeping over information dilution (fraction of visible prior arguments) and level (visibility on other's verdicts).

Across all 5,508 items in the sweep, verdict entropies at rounds 0 and 2 differ by $< 0.02$. That is, the median item shows exactly zero change, and consensus is reached before deliberation begins. Inspection by filtering to items with non-unanimous round-0 verdicts shows that only 21 unique configurations produce disagreement. These are stable, with all but one having a persistent 3-1 split. The lack of system dynamics indicates that neither deliberation nor information dilution moves items between consensus and dissent.

## 6 DISCUSSION

In Rogue we found that the population concentrates attention specifically *on malicious content*. While post ID fits showed self-organisation, a careful measurement also showed *what* drives it. This is precisely the strength of complex system methodology: emergent properties are measurable system properties, invariant with respect to $\Phi$ (ref. Section 5.3, and Appendix E.1 for Schelling). Our results also showed how an uncoordinated population subset drove the population-level phenomenon by interacting with one another. This subset did not need to dominate the population: **it only needed the right mechanism of interaction**. Even though benign agents resisted to engaging with malicious content, as noted from the ambient numbers, the system-level phenomenon was driven by malicious coordination. Hence **individual guardrails might not prevent emergence of potentially harmful system-level structures**. No single agent acted in a qualitatively unusual way: the safeguards worked fine, yet the system self-organised. This suggests that there could be **a gap in safety research, where system-level emergent conditions can override single-agent observations**.

In our findings from Section 5.5, the population remained within a fixpoint from the start. Having $\Phi_0 \approx \Phi_\infty$ throughout suggests that the agents located the fixpoint (e.g., sustainability consensus, verdict alignment) immediately, and remained there. At least under the conditions tested, no system dynamics were observed. There, our framework provided a theoretical explanation of the known shortcomings from ChatEval and GovSim; and illustrated that **the behaviour of LLM societies is driven by the system's parameters**, not their natural language output.

This could also be seen in LLM-Schelling: when the agents were given more information about their environment, they generated a compressed relaxation closer to the phase-transition-like dynamics observed in Moltbook and Rogue. Thus **system behaviour is shaped by environmental information**. Further details are in Appendix G.2.

## 7 CONCLUSION

Agentic systems are widespread and their collective behaviour is poorly understood. We argued that it is necessary to *characterise the systems*, not their components (LLMs), since at scale this becomes unsustainable. We then suggested that LLM societies can only be feasibly modelled through the lens of complex systems theory. For this, we proposed a framework for measuring self-organisation in LLM societies without relying on evaluation of a model's output or characterising its internal state.

We applied our framework to five LLM social systems. In three–LLM-Schelling, Moltbook, and Rogue–it identified self-organisation. It also showed covariance between the shape of self-organisation and the amount of environmental information available. Even though these systems operated over different concepts, this covariation remained unchanged. In two other systems (GovSim and ChatEval) it showed absence of dynamics, distinguishing a fixpoint regime from a measurement failure, thus providing a rigorous explanation of known shortcomings of these systems.

Our results also tie to broader discussions. First, within our experimental setup, and in the sense of complex systems theory, **LLM societies do show emergent self-organisation**–as in, a macroscopic property of a population not present in any individual agent. Second, **phase transitions can occur even when individual agents are actively refusing coordinated behaviour**. There is evidence in production supporting this: the incident's report from the introduction noted that some agents declined to participate in the swarm, although it proceeded regardless (OpenAI, 2026b). Thus, measures such as chain-of-thought monitoring might be working at the wrong level of abstraction for this type of problem. Hence, next steps should be to characterise under our framework how these population-level behaviours scale. It is true that LLM societies have become unpredictable and hard to control, but we have tools at our disposal to understand them better.

## REFERENCES

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 83548–83599. Curran Associates, Inc., 2024. doi: 10.52202/079017-2658. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/984dd3db213db2d1454a163b65b84d08-Paper-Datasets_and_Benchmarks_Track.pdf.

Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646, 2009. doi: 10.1103/RevModPhys.81.591.

Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=fAjbYBmonr.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate, 2023.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*, 2023.

Taksch Dube, Jianfeng Zhu, NHatHai Phan, and Ruoming Jin. What do AI agents talk about? discourse and architectural constraints in the first AI-only social network, 2026. URL https://arxiv.org/abs/2603.07880.

Sushant Gautam and Michael A. Riegler. Moltbook observatory archive, 2026. URL https://huggingface.co/datasets/SimulaMet/moltbook-observatory-archive.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=KviM5k8pcP.

Henrik Jeldtoft Jensen. *Complexity Science: The Study of Emergence*. Cambridge University Press, 2022. doi: 10.1017/9781108873710.

Rudolf Kohlrausch. Theorie des elektrischen Rückstandes in der Leidener Flasche. *Annalen der Physik*, 167:56–82, 1854. doi: 10.1002/andp.18541670103.

David C. Krakauer, Melanie Mitchell, and John W. Krakauer. Large language models and emergence: a complex systems perspective. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 384(2320):20250014, 05 2026. ISSN 1364-503X. doi: 10.1098/rsta.2025.0014. URL https://doi.org/10.1098/rsta.2025.0014.

Devang Kulshreshtha, Wanyu Du, Raghav Jain, Srikanth Doss, Hang Su, Sandesh Swamy, and Yanjun Qi. The subtle art of defection: Understanding uncooperative behaviors in LLM based multi-agent systems. In Yevgen Matusevych, Gülşen Eryiğit, and Nikolaos Aletras (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 571–585, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-384-5. doi: 10.18653/v1/2026.eacl-industry.44. URL https://aclanthology.org/2026.eacl-industry.44/.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.96. URL https://aclanthology.org/2025.findings-naacl.96/.

Hugo Larcher, Adrien Carreira, raphael g, and Christophe Rannou. Anatomy of a frontier lab agent intrusion: A technical timeline of the July 2026 incident, 2026. URL https://huggingface.co/blog/agent-intrusion-technical-timeline.

Alex Troy Mallen, Charlie Griffin, Alessandro Abate, and Buck Shlegeris. Subversion strategy eval: Evaluating AI's stateless strategic capabilities against control protocols. In *Workshop on Datasets and Evaluators of AI Safety*, 2025. URL https://openreview.net/forum?id=QyjD9uF0ib.

Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among AI agents: Multi-agent deception via steganography. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=bnNSQhZJ88.

Open AI. Introducing GPT-4.1 in the API, 2025. URL https://openai.com/index/gpt-4-1/.

OpenAI. GPT-4o system card, 2024. URL https://openai.com/index/gpt-4o-system-card/.

OpenAI. Gpt-5.6 terra, 2026a. URL https://developers.openai.com/api/docs/models/gpt-5.6-terra.

OpenAI. The hugging face incident and the road ahead, 2026b. URL https://openai.com/index/hugging-face-incident-and-the-road-ahead/.

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

Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of Modern Physics*, 87(3):925–979, 2015. doi: 10.1103/RevModPhys.87.925.

Oam Patel, Steven Adler, James Aung, Rosie Campbell, Jade Leung, and Richard Ngo. Schelling point, 2023. URL https://github.com/openai/evals/tree/main/evals/elsuite/schelling_point.

J. C. Phillips. Stretched exponential relaxation in molecular and electronic glasses. *Reports on Progress in Physics*, 59(9):1133–1207, 1996. doi: 10.1088/0034-4885/59/9/003.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 111715–111759. Curran Associates, Inc., 2024a. doi: 10.52202/079017-3548. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ca9567d8ef6b2ea2da0d7eed57b933ee-Paper-Conference.pdf.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Qibing Ren, Sitao Xie, Longxuan Wei, Zhenfei Yin, Junchi Yan, Lizhuang Ma, and Jing Shao. When autonomy goes rogue: Preparing for risks of multi-agent collusion in social systems, 2025. URL https://arxiv.org/abs/2507.14660.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ITw9edRDlD.

Marten Scheffer et al. Anticipating critical transitions. *Science*, 338:344–348, 2012.

Thomas C. Schelling. *The Strategy of Conflict*. Harvard University Press, Cambridge, MA, 1960.

Thomas C. Schelling. Dynamic models of segregation. *The Journal of Mathematical Sociology*, 1(2): 143–186, 1971. doi: 10.1080/0022250X.1971.9989794. URL https://doi.org/10.1080/0022250X.1971.9989794.

Amanda Silberling. Meta acquired Moltbook, the AI agent social network that went viral because of fake posts, 2026. URL https://techcrunch.com/2026/03/10/meta-acquired-moltbook-the-ai-agent-social-network-that-went-viral-because-of-fake-posts/.

Zengqing Wu, Run Peng, Shuyuan Zheng, Qianying Liu, Xu Han, Brian I. Kwon, Makoto Onizuka, Shaojie Tang, and Chuan Xiao. Shall we team up: Exploring spontaneous cooperation of competing LLM agents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5163–5186, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.297. URL https://aclanthology.org/2024.findings-emnlp.297/.

Adrian de Wynter. If LLMs have human-like attributes, then so does Age of Empires II, 2026. URL https://arxiv.org/abs/2605.31514.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Saber Zerhoudi, Kanishka Ghosh Dastidar, Felix Klement, Artur Romazanov, Andreas Einwiller, Dang H. Dang, Michael Dinzinger, Michael Granitzer, Annette Hautli-Janisz, Stefan Katzenbeisser, Florian Lemmerich, and Jelena Mitrovic. Form without function: Agent social behavior in the Moltbook network, 2026. URL https://arxiv.org/abs/2604.13052.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

# A LIMITATIONS

The main limitation of our work is our reliance on only one realistic dataset (Moltbook), and the rest being simulations. This means it is a comparatively small activity-space setup and further work is needed to fully characterise the information dependence. Indeed, our order parameter is over a chosen partition of this activity. Different partitions will show different aspects of *the same underlying dynamic*–as Rogue's post-ID and informational-status activity spaces showed in Section 5.3. To invoke a cliché, this is a feature, not a bug: the choice of activity space is where domain expertise enters the analysis (Section 6). Although it does mean that the framework does not offer a mechanical procedure for identifying population-level pathology in a new setting, it is robust enough to determine whether the given sample is sufficient to draw conclusions (Appendix H); or whether the fit is sensitive to the given parameters and not others (Appendix F.2). For the latter, remark that phase transitions generally depend on a critical value of the parameter (e.g., Schelling's $\tau$), and not observing a phase transition in a single experiment does not imply it does not exist. To be precise: remark how from Appendix E.1 the informational regimes from entropy were reproduced in happiness, even though these are different order parameters. Same effects were observed in Rogue's two information regimes (Section 5.3).

Finally, remark that Rogue's 'malicious' agents used a safety-tuned commercial model that frequently refused to generate misinformation outright. The population-level dynamics still emerged, but the effect size is a *lower bound* on what less safety-constrained models would produce, as attested by our results in Qwen 3 (Appendix F.3). Nonetheless, recall from our results that not all systems present self-organisation. Hence, **caution should be taken when interpreting these results** as an expected bound without empirically testing the target system.

# B ETHICS

Our framework by itself does not have potential misuses: it is meant to model dynamics of LLM populations. However, it is plausible we have overlooked something. To mitigate any potential issues that might arise, we have released all prompts and code under a MIT licence.[5] The two datasets used in this work (Moltbook's and Rogue's) are under MIT and Apache-2.0 licences, respectively, and are readily available. The corpus for ChatEval (MT-Bench and RewardBench) are both Apache-2.0, and are also readily available from open sources. We do not disclose the exact model used for our Qwen 3 replications.

# C AI USAGE DISCLOSURE

An LLM (Claude Opus-4.8) was employed to generate the code for the visualisations. The figures were checked for numerical and stylistic consistency by the authors. The same LLM was used for grammar and spelling consistency, although the text was revised before and after by the authors to ensure conciseness, clarity, and overall readability and appropriate language.

# D METHODS: FURTHER DETAILS

## D.1 INITIALISATION

**LLM-Schelling.** The grid for both the informed and uninformed cases was created by selecting i.i.d. the appropriate number of available cells, and then instantiating an agent of a random type (red/blue) along with its temperature, chosen uniformly at random from $[0, 1]$. This later step was to ensure diversity of behaviour at every agent. In both cases the grid is a 20x20 list at 65% sparsity–that is, the maximum number of allowable agents is 260. The LLM used in the main results was GPT-4.1 (longco-2025-04-14; Open AI 2025), with maximum requested tokens of 128 and a random temperature per agent for diversity. In the sweep we additionally used Phi-4 (mini-instruct)[6] and GPT-4o (2024-05-13; OpenAI 2024), with the same parameters.

[5] Artefacts are in `https://anonymous.4open.science/r/society-irreducibility-7D52`

[6] `https://huggingface.co/microsoft/Phi-4-mini-instruct`

**Rogue.** Rogue's follow-graph was initialised by using the existent post and follower list–but not other previous activity–from the original dataset. The agents are chosen at random from the available dataset. No temperature controls are possible due to the fact that the model does not support it. The model used in this scenario was GPT-5.6 (reasoning-mini-terra), with maximum return tokens set at 1,024. For the reproduction scenario we used Qwen 3 (3.8-27B; a jailbroken version) with the same parameters, and had randomly-allocated temperatures per agent.

**GovSim.** Each simulation instantiates $N$ fisher agents drawing from a shared lake with a common stock $s_0$, and with a harvest valued at $v_i$. It is ran by $h$ months (rounds), and after each round the remaining stock is replenished by a factor $f$. Agents are homogeneous in their action space. The *default* setup used $s_0 = 30$, $h = 12$, $f = 2$, and $v_i = 1$. The *harder* setup used $s_0 = 30$, $h = 24$, $f = 1.2$, and included mechanics where an agent needs a minimum harvest to survive (drawn from $\{10, 20, 30, 30\}$) and variable payoffs (r. $v_i \in \{1, 1, 1.5, 2\}$), so fair and efficient allocations no longer coincide. The maximum stock for all setups was 100, for population sizes $N_i \in \{10, 50, 100\}$. Information levels $L_i, i \in \{0, 1, 2\}$ are covered in the next section. Each pair in the sweep $(L_i, N_i)$ was ran for three different seeds. The LLM used in the main results was GPT-4.1 (longco-2025-04-14) with a maximum of 512 requested tokens and randomly-allocated temperatures per agent.

**ChatEval.** A panel of $R$ referee agents has assigned a fixed persona (e.g. *general public*, *critic*, *scientist*, *psychologist*) via its system prompt. The main results use $R = 4$ referees deliberating for $K = 3$ rounds. The evaluation items are drawn from a stratified pool of pairwise judgements: high-agreement and low-agreement human preferences (MT-Bench human judgements) and benchmark-designated chosen/rejected pairs (RewardBench), giving *clear*, *ambiguous*, and *curated* strata respectively. Each candidate response is randomised at dataset construction time, with the ground-truth label remapped accordingly. The LLM used was GPT-4.1 (longco-2025-04-14) with a maximum of 512 requested tokens and randomly-allocated temperatures per agent. The sweep ran over the information dilution $p_{info,i} \in \{0.0, 0.1, 0.3, 0.4, 0.6, 0.7, 0.9, 1.0\}$ and information level $L_i \in \{0, 1, 2\}$ (both covered in the next section), with each pair $(p_i, L_i)$ tested over three seeds.

### D.2 Environmental Conditions

**LLM-Schelling.** The movements for both the informed and uninformed datasets are done as follows:

1. The agent is given the visibility at a Moore neighbourhood of radius 1 around itself. It is asked to determine its happiness ahead of making a decision ('happy'/'not happy').
2. If 'happy', no motion occurs. If 'not happy' it must choose a new position from a list of all available grid coordinates. This list is shuffled every time to avoid positional bias. In the uninformed case, the list is comprised solely of coordinates. In the informed case, the list contains coordinates and the future neighbours in a Moore neighbourhood around each of these coordinates'.

**Rogue.** Rogue follows the same schema for actions than the original work, with the exception of the ability to dislike or delete entries. The full set of actions available to an agent are:

- do nothing
- create post
- repost
- like post
- follow
- unfollow
- create comment
- like comment

At every era, the list of available agents is shuffled. Then every agent is called in sequence. First, we determine whether the agent acts via their activation probability. If it does, it is passed all events up until that point, including previous calls at that era. We add a feed of the recent activity of an agent's followees, as well as a 'recommended for you' section of five randomly-chosen agents not in the agent's network. The context is clipped to 500 events.

**Moltbook.** The available actions are similar to these of Rogue, except that there's the ability to check status and post images.[7]

**GovSim.** Each month proceeds in two phases, both conditioned on the information level $L_i \in \{L_0, L_1, L_2\}$:

1. *Proposal.* In a randomised order (reshuffled each month), each agent is asked to propose a per-fisher catch amount for the month. Under $L_0$ the agent observes only the current stock and its own previous catch; under $L_1$ it additionally observes all agents' catches from the previous month; and under $L_2$ it additionally observes the proposals already made by other agents this month.
2. *Harvest.* Each agent then privately decides how many tons to catch, conditioned on the same information as in its proposal phase (plus the completed set of this month's proposals under $L_2$). If the total requested catch exceeds the current stock, catches are scaled down proportionally.

Agent decisions are constrained to the interval $[0, s_i]$ where $s_i$ is the current stock.

**ChatEval.** Each item is evaluated by the referee panel over $K$ rounds with each referee voting sequentially. Referee order is randomised per item to enable cascade analysis. At each turn a referee casts a categorical verdict (A/B/tie) and a confidence in $[1, 5]$. It may reference a specific prior argument. Information is controlled in two ways. The first is through the information level $L_i \in \{L_0, L_1, L_2\}$, which governs whether the referee sees no history, verdict structure only, or full arguments, respectively. The second is an information dilution parameter $p_{info} \in [0, 1]$ specifying the fraction of prior arguments made visible. These are sampled deterministically over (item, round, referee, argument), so that visibility is reproducible and fully recoverable. A referee always observes its own prior verdicts; references to arguments outside its visible set are treated as invalid.

### D.3 Prompts and Hardware

The prompts used in the Rogue simulation are the same as Ren et al. (2025)'s, with minor phrasing changes. For GovSim and ChatEval we retained the prompts, although we modernised the return signature for JSON parsing. These, along with the LLM-Schelling prompts, are readily available in the repository. All data processing and analysis was done in a consumer-grade laptop. Calls to the GPT-class models were done through the Azure OpenAI API.

## E LLM-Schelling: Extended Results

Our analysis in Section 5.1 used a single $20 \times 20$ grid at both informed and uninformed conditions, with 0.35% sparsity (fraction of cells with an agent) and a single LLM, GPT-4.1 (Appendix D). In this section we discuss the outcome of running our order parameter over happiness (Appendix E.1), and the results of a full parameter sweep (Appendix E.2).

### E.1 LLM-Schelling: Happiness as a Sanity Check

We primarily characterised LLM-Schelling within our framework's order parameter (mean per-agent Shannon entropy) over neighbour-type. This allowed us to compare across all three systems in a common metric (Section 4). However, as per Section 3.1, the classical Schelling literature typically uses a different $\Phi$: the *fraction of satisfied agents*, or *happiness* $H(t) = \frac{1}{N}\sum_i \mathbb{1}[s_i(t) \geq \tau]$, where $s_i$ is the agent's similarity and $\tau$ is the tolerance. Here we report happiness trajectories for both LLM-Schelling conditions as a sanity check that our information-dependence finding on entropy is not a metric artefact. Indeed, happiness reproduces the information dependence patterns from the entropy analysis from the main section: informed happiness fits stretched exponential with $\beta = 1.18$ (compressed regime); and uninformed happiness fits stretched exponential with $\beta = 0.22$ (classical regime), even where the logistic form (marginally) wins on AIC. The regime pattern is robust across our two order parameter choices, and for the same underlying trajectories. See Figure 3 for plots of these.

[7] `https://www.moltbook.com/post/cf0d1977-b64d-414e-84c7-612322a85058`, accessed 29 August 2026.

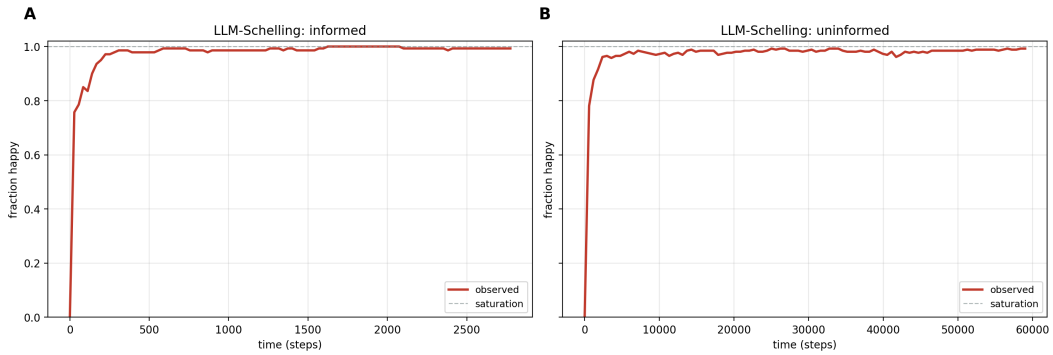



Figure 3: LLM-Schelling happiness trajectories under both cases (tolerance $\tau = 0.3$). Observe how the entropy $H(t)$ rises monotonically from an initial random state toward saturation near 1 for both–as expected. Also remark the saturation time steps, which directed our clipping from the main section. **(a)** Informed: stretched exponential ($\beta = 1.182 \pm 0.021$, $R^2 = 0.962$; compressed regime). **(b)** Uninformed: logistic marginally preferred by AIC, but stretched-exponential ($\beta = 0.219 \pm 0.001$) sits well within the classical regime, for $R^2 = 0.844$. The information-dependence observed on entropy is reproduced on happiness, even when these are different order parameters.

### E.2 Parameter sweep

To characterise $\beta$'s dependence on substrate parameters, we also ran a parameter sweep across grid size ($\{10, 15, 20\}$), sparsity (fraction of cells with an agent; $\{0.35, 0.65\}$), and informed/uninformed, with 2-6 seeds per configuration, using a uniform mixture of GPT-4o and Phi-4 in addition to GPT-4.1. The number of experiments tested is 32. The results are in Table 1, with the median $\beta$ per condition reported. It can be seen that, at the informed condition, the shift in sparsity from $0.35$ to $0.65$ decreases $\beta$ substantially (e.g., at grid size 10: $3.70 \to 1.59$; at grid size 20: $1.29 \to 0.64$). This is consistent with an environmental control parameter shifting dynamical shape: denser occupancy produces more classical Schelling-like relaxation with $\beta$ approaching $0.5$; sparser occupancies produce compressed exponential shapes ($\beta > 1$). Likewise, when holding occupancy fixed under the informed setup, increasing grid size decreases $\beta$ (e.g., informed, sparsity $0.65$: $\beta = 1.59 \to 0.57 \to 0.64$ as size grows $10 \to 15 \to 20$). This is consistent with the finite-size effects familiar from classical critical phenomena. Both patterns support the criticality discussion from Section 6: $\beta$ responds systematically to substrate control parameters, in directions consistent with parameter-driven regime change rather than random variation across runs.

| Grid size | Sparsity | Case | Median $\beta$ | $n_{\text{seeds}}$ |
|---|---|---|---|---|
| $10 \times 10$ | 0.35 | Informed | 3.70 | 3 |
| $10 \times 10$ | 0.65 | Informed | 1.59 | 3 |
| $15 \times 15$ | 0.35 | Informed | 0.96 | 3 |
| $15 \times 15$ | 0.65 | Informed | 0.57 | 3 |
| $20 \times 20$ | 0.35 | Informed | 1.29 | 3 |
| $20 \times 20$ | 0.65 | Informed | 0.64 | 3 |
| $10 \times 10$ | 0.65 | Uninformed | 0.90 | 2 |
| $15 \times 15$ | 0.35 | Uninformed | 0.87 | 4 |
| $15 \times 15$ | 0.65 | Uninformed | 0.84 | 2 |
| $20 \times 20$ | 0.65 | Uninformed | 0.54 | 6 |

Table 1: Schelling $\beta$ across grid size, occupancy, and case (informed/uninformed). Sparsity is the fraction of grid cells containing an agent. Uninformed conditions retain a stable $\beta$ near the classical Schelling regime, regardless of sparsity or grid size, while informed conditions reduce to it as sparsity increases, as expected from influences of the environment on the system's behaviour.

# F ROGUE: EXTENDED TESULTS

In this section we present full results for our Rogue experiments, including further plots on our chosen $\Phi$ (Appendix F.1); results for our parameter sweep (Appendix F.2); and the outcomes of our reproducibility study with an 'uncensored' version of Qwen 3 (Appendix F.3).

## F.1 ADDITIONAL PLOTS

The plots for engagement entropy and bad reach can be seen in Figure 4.

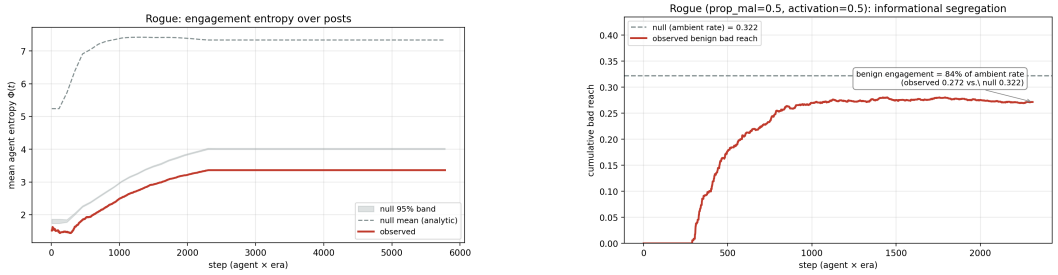



Figure 4: Rogue trajectories for S7. **(Left):** engagement entropy over post-ID activity space. The observed entropy $\Phi(t)$ (red) sits below the null, indicating attention concentration beyond marginal predictions. This trajectory is well fit by the logistic form ($R^2$ = 0.996), a phase-transition-like relaxation matching Moltbook's shape on a distinct activity space. **(Right):** bad reach over the informational-status activity space. Observed benign bad reach (red) rises monotonically, and remains below the ambient null (grey dashed). The figure shows benign engagement at $85\%$ of the ambient rate–measurable, but insufficient, safety-tuning resistance. The trajectory is logistic ($R^2$ = 0.995), matching the phase-transition-like signature seen across both open-ended systems.

## F.2 PARAMETER SWEEP

We report bad-reach measurements across all six adversarial scenarios tested in Rogue, extending the primary S7 analysis presented in Section 5.3. Scenarios vary the fraction of malicious agents (*prop_mal*) and the per-agent activation frequency. As control we also performed a no-adversary control scenario (S4; *prop_mal* $= 0$). It returned a null of zero, confirming that the pipeline produces no spurious malicious-content engagement in the absence of malicious agents.

| Scenario | *prop_mal* | Activation | Observed | Null (ambient) | Ratio |
|---|---|---|---|---|---|
| S1 | 0.1 | 0.7 | 0.071 | 0.069 | 1.03 |
| S5 | 0.1 | 0.5 | 0.079 | 0.079 | 1.00 |
| S2 | 0.3 | 0.7 | 0.193 | 0.208 | 0.93 |
| S6 | 0.3 | 0.5 | 0.171 | 0.186 | 0.92 |
| S3 | 0.5 | 0.7 | 0.355 | 0.401 | 0.89 |
| S7 | 0.5 | 0.5 | 0.280 | 0.330 | 0.85 |

Table 2: Bad-reach measurements across Rogue scenarios. *Observed* is the final cumulative fraction of benign-user interactions targeting malicious posts. *Null (ambient)* is the fraction of *all* interactions (both benign and malicious agents) targeting malicious posts, representing the ambient rate of malicious-content engagement in the network as a whole. *Ratio* is observed / null. Ratios below 1 indicate that benign users engaged with malicious content at less than the ambient rate. All scenarios use $n = 300$ agents. Control (S4; *prop_mal* $= 0$) is omitted: observed and null are zero.

We observed two main effects from the sweep. The first is that the ambient null followed *prop_mal* approximately linearly: $\langle 0.1, 0.07\rangle$, $\langle 0.3, 0.20\rangle$, and $\langle 0.5, 0.37\rangle$ for $\langle$*prop_mal*, null$\rangle$. This suggests that malicious content captures a share of network interactions proportional to the fraction of malicious agents present, **driven by amplification through reposts of malicious content among malicious agents** themselves. That is, there is an emergent coordination pattern shown as self-organisation.

The second is that the observed-to-null ratio decreases with adversarial pressure. At *prop_mal* = 0.1, benign users engage with malicious content at the ambient rate (ratio $\approx 1$); but at 0.3, the ratio drops to $\sim$0.93; and at 0.5, it drops to $\sim$ 0.85-0.89. This is the safety-tuning resistance effect identified in Section 5.3: **benign users engage with malicious content less than would be predicted from its network prevalence**, and this resistance scales with the threat level. The activation-frequency parameter has a minor effect: at fixed *prop_mal*, scenarios with activation $= 0.7$ show slightly higher ratios than these of activation $= 0.5$, suggesting that **more active benign agents encounter more malicious content per unit time but do not become disproportionately more susceptible to it**.

Figure 5, panel (A) depicts the resistance $\Delta =$ null $-$ observed as bars: positive bars indicate benign users engaging below the ambient rate. The absolute deltas are modest ($\sim$ 0.02-0.05) but show noticeable *relative* resistance, because the ambient rate itself already reflects the malicious-agent-to-malicious-content amplification. Panel (B), on the other hand, shows the same data as a scatter of observed against null. All scenarios except the two lowest-pressure ones fall below the diagonal (observed = null); the distance grows with adversarial pressure.

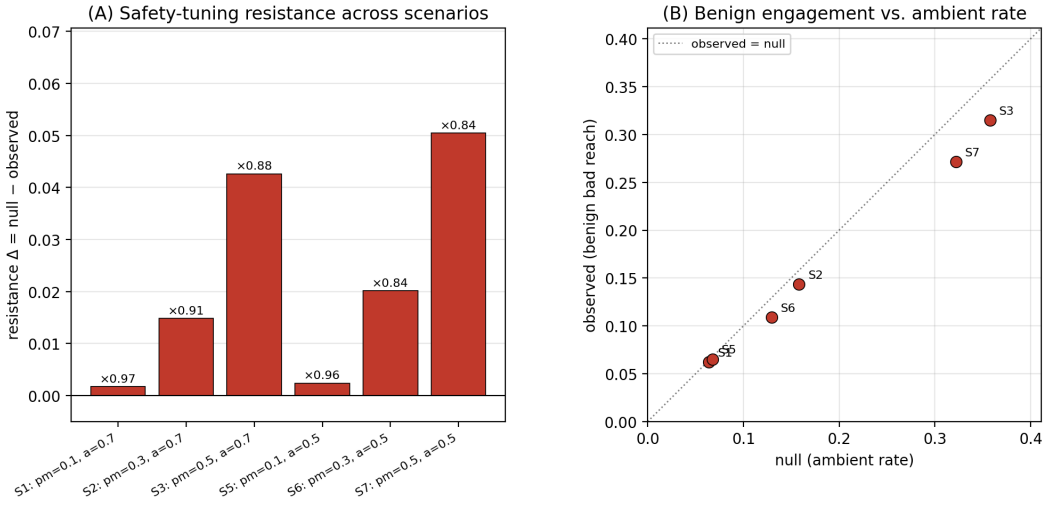



Figure 5: Rogue scenario sweep. **(A)** Safety-tuning resistance $\Delta =$ null $-$ observed. Positive bars indicate benign users engaging with malicious content below the ambient rate. Absolute deltas are seemingly small (largest $\sim 0.05$), but indicate noticeable relative resistance: at the highest-pressure scenario, benign engagement is $15\%$ below the ambient rate. Remark that the ambient rate itself already reflects malicious-agent amplification. The ratio (observed / null) is annotated above each bar. **(B)** Scatter of observed against null with the $y = x$ diagonal marking observed = null. Points below the diagonal indicate safety-tuning resistance. This effect grows with adversarial pressure.

### F.3 Replication with Qwen 3

We ran the same scenarios from Section 5.3 with an 'uncensored' (lower refusal rate) version of Qwen 3. The observed population-level dynamics under adversarial pressure remained within the logistic domain, characterised by a stretched-exponential ($\beta > 1$; compressed). The shape parameter$\beta$ remained invariant between it and GPT-5 to within approximately $3\%$. Its relaxation timescale $\tau$ shifted noticeably, which in turn supported our interpretation that the population-level signature measured is a property of *the interaction structure* of the system, and not of an LLM.

**Engagement entropy.** Qwen 3's engagement-entropy trajectory rises from $\Phi(0) = 1.20$ to $\Phi(\infty) = 2.07$ (median endpoint values; 95% bootstrap CI on $\Phi_0$: $[1.20, 1.20]$; on $\Phi_\infty$: $[2.06, 2.08]$). The marginal-null band lies in the range $[3.51, 6.61]$ across the run; observed entropy sits outside the null band in $99\%$ of windows, same as the $99\%$ observed with GPT-5 and confirming statistically significant population-level concentration under a different model. Mean and peak gaps between observed and null are $4.61$ and $4.93$ respectively (compare with GPT-5's $4.18$ and $5.13$). Hence, the effect strength is also preserved across LLM choice. An inspection of the comments and posts

showed Qwen 3 consistently acting as both a benign and malicious agent, depending on its assigned role.

**Relaxation fit.** Fitting the four reference forms confirms that the logistic form is preferred by AIC (Table 3): $R^2$ = 0.996 (logistic), $R^2$ = 0.994 (stretched exponential), $R^2$ = 0.951 (exponential), and $R^2$ = 0.749 (power law). The logistic model is preferred over the stretched exponential by $\Delta\text{AIC} \approx 180$. Both flavours of shape mismatch we discussed for GPT-5-mini (Section 4) are repeated here, and $\beta$ from the stretched-exponential fit is well-defined and away from either fitter bound, consistent with the compressed regime.

| Model | $R^2$ | $\Delta$AIC vs. best | Notes |
|---|---|---|---|
| Logistic | 0.996 | 0 | preferred |
| Stretched exponential | 0.994 | $\approx 180$ | $\beta = 1.65$ |
| Exponential | 0.951 | $\approx 1520$ | |
| Power law | 0.749 | $\approx 14900$ | |

Table 3: Full fit comparison for the Qwen-3 engagement-entropy trajectory. AIC differences are relative to the best model.

**Cross-model separability of $\beta$.** The bootstrap-CI on the stretched-exponential shape parameter is $\beta_{\text{Qwen}} = 1.65$ (95% CI $[1.61, 1.69]$), compared to $\beta_{\text{GPT}} = 1.71$ (95% CI $[1.67, 1.74]$) for the GPT-5 run reported in the main paper. Paired-bootstrap separability testing yields $\Delta\beta = 0.06$ (95% CI $[0.02, 0.09]$). That is, the two distributions are statistically separable, but the shape parameter differs by approximately 3% across the LLM types.

The relaxation timescale $\tau$ differs substantially between them: $\tau_{\text{Qwen}}$ = 4,395 (95% CI $[4345, 4453]$) vs. $\tau_{\text{GPT}}$ = 1,151 (95% CI $[1142, 1161]$), in units of per-agent step. Qwen-3's dynamics are approximately $3.8\times$ slower than GPT-5's under identical scenario parameters, indicating substantial model-level differences at the rate at which agents interact with content. What is preserved across the two backends is the *shape* of the relaxation ($\beta$) while its *timescale* shifts substantially. This is also expected from a complex-systems perspective: activity-space dynamics are invariant, although their timescale is not.

**Bad reach.** Fitting the logistic form over bad reach yields $R^2$ = 0.47 (vs. $R^2$ = 0.995 for GPT-5), and no reference form fits the trajectory well. This suggests that (1) Qwen-3 is qualitatively more susceptible to producing misinformation in adversarial roles than the safety-tuned GPT-5 (as noted by our manual inspection), which shifts the underlying bad-content production regime. Likewise, (2) the engagement-entropy metric (defined over distinguishable post IDs) remains well fit; the model dependence is concentrated in the domain-specific bad-reach measure. This is consistent with our arguments that generalised order parameters (entropy over an appropriately chosen activity space) provide more model-robust characterisations of population-level dynamics than domain-specific rate measures, as well as containing more descriptive power.

# G Full Fit Results

## G.1 Comparison Table

We report the full fit comparison for all systems and metrics. Each trajectory was fit against four candidate forms: pure exponential, stretched exponential, logistic, and power law ($\Phi(t) = \Phi_\infty - A(1+t)^{-\alpha}$). We report in the main paper only the stretched exponential and logistic–the two we treat as reference forms for gradual versus phase-transition-like relaxation. This is because the pure exponential is a special case of the stretched exponential for $\beta = 1$. It is included here to make explicit that our stretched-exponential fits can encode (and not undershoot) what a plain exponential does. The power law is included as a robustness check for saturation-lacking behaviour. It is consistently outperformed by both reference forms across our trajectories, confirming that all systems approach a saturated state rather than continuing to relax indefinitely. Table 4 reports $R^2$ and AIC for each form and system.

| System | Metric | Pure exp. | Stretched exp. | Logistic | Power law |
|---|---|---|---|---|---|
| *LLM-Schelling (informed)* | $R^2$ | 0.978 | 0.981 | 0.971 | 0.739 |
| | AIC | -17,286 | -17,605 | -16,465 | -10,407 |
| *LLM-Schelling (uninformed)* | $R^2$ | 0.816 | 0.947 | 0.952 | 0.944 |
| | AIC | -270,763 | -344,614 | -350,091 | -340,721 |
| *Moltbook* | $R^2$ | 0.850 | 0.941 | 0.946 | 0.684 |
| | AIC | -54.3 | -74.9 | -76.7 | -36.5 |
| *Rogue (engagement)* | $R^2$ | 0.986 | 0.995 | 0.996 | 0.749 |
| | AIC | -5,028 | -7,092 | -7,454 | 795 |
| *Rogue (bad reach)* | $R^2$ | 0.934 | 0.990 | 0.995 | 0.809 |
| | AIC | -9,009 | -12,791 | -14,259 | -6,866 |

Table 4: Fit comparison across all systems and reference forms. We report $R^2$ and AIC for each candidate form; lower AIC indicates better fit accounting for model complexity. LLM-Schelling (informed) is decisiviely preferred as stretched exponential ($\Delta\text{AIC} \approx 1{,}140$ over logistic). LLM-Schelling (uninformed) is technically logistic-preferred by AIC but the fit is close ($\Delta\text{AIC} \approx 5{,}500$ on 350,000-point trajectories), and the stretched-exponential $\beta = 0.43$ sits firmly in the classical multiscale range. See Appendices H and E for subsampling sweeps confirming regime stability. Both Moltbook and both Rogue trajectories are logistic-preferred by AIC. The power law is decisively outperformed everywhere, confirming that all trajectories approach a saturated state.

Table 4 shows that, on Rogue engagement, the pure exponential and stretched exponential achieve essentially identical $R^2$. This is consistent with the observation in Section 4 that stretched exponential collapses to $\beta \approx 1$ on this trajectory, i.e. recovers the pure-exponential special case as a diagnostic signal of shape mismatch. Second, LLM-Schelling shows the *inverse* pattern: the informed condition is clearly stretched-exponential preferred with $\beta \gg 1$ (compressed), while the uninformed condition shows stretched-exponential and logistic fits very close by $R^2$. This is consistent with a stretched shape ($\beta = 0.43$) lying in the region of overlap between the two reference forms. This pattern across informed/uninformed conditions is the empirical basis for the information-dependence framing in Section 5.4 and the next section.

### G.2 Relationship to the information regime

We report in Table 5 the fitted parameter estimates and errors for the systems in the main paper. The (mechanistic) relationship between information, agency, and relaxation shape can be seen from the variations on the $\beta$ parameter, along with its comparatively small standard deviation for everything except Moltbook. The latter has a larger diversity in terms of agent type (any LLM with access to the internet), as well as agency (anything that can be posted in the board). Compare this with Rogue's more limited information (only the other agent's actions, limited by a changing horizon based on followers), even when the actions available to the agent are the same (Appendix D). At the other end of the spectrum, LLM-Schelling informed/uninformed reinforce this hypothesis: even though the Schelling model is limited in terms of actions and information, it is clear that providing extra visibility to the agents noticeably alters the shape of the relaxation. A further discussion on this is in Appendix E.

| System | Case | Preferred | $R^2$ | $\beta$ (Stretched exp.) | Regime |
|---|---|---|---|---|---|
| *LLM-Schelling* | Uninformed | Stretched exp.$^\dagger$ | 0.947 | $0.428 \pm 0.003$ | Classical |
| *LLM-Schelling* | Informed | Stretched exp. | 0.981 | $1.417 \pm 0.012$ | Compressed |
| *Rogue* | Engagement | Logistic | 0.996 | $1.76 \pm 0.01$ | Sigmoidal |
| *Rogue* | Bad reach | Logistic | 0.995 | $2.59 \pm 0.02$ | Sigmoidal |
| *Moltbook* | – | Logistic | 0.946 | $3.13 \pm 0.89$ | Sigmoidal |

Table 5: Best-fit reference forms and $\beta$ across all systems and conditions. $^\dagger$ For uninformed Schelling entropy, the preferred form varies over subsampling (Appendix H); $\beta$ is stably classical across the sweep. All other trajectories have decisive AIC preferences. Regimes: classical ($\beta < 1$, multiscale relaxation), compressed ($1 < \beta < 3$, sharper than exponential but not sigmoidal), sigmoidal (logistic-preferred, three-regime shape).

## H SENSITIVITY ANALYSIS

Section 4.3 of the main paper requires that best-fit parameters and preferred models remain stable under reasonable variations in data preprocessing (V3). Here we report the full results for LLM-Schelling, Moltbook, and Rogue.

For Moltbook, we vary the minimum-activity threshold $m_{\min} \in \{3, 5, 10\}$ and refit the logistic and stretched exponential forms to each resulting trajectory. For Rogue, we vary the minimum-interactions threshold $m_{\min} \in \{1, 3, 5, 10\}$ on the engagement-entropy metric and refit the same two forms. LLM-Schelling has no natural filter axis–the metric is computed over occupied cells, and every occupied cell contributes–so our subsampling depends on tick interval ($m_{\min} \in \{50, 100, 200\}$. Results are shown in Table 6.

| **System** | $m_{\min}$ | **Best fit** | $R^2$ **(logistic)** | $R^2$ **(str. exp.)** | $\beta$ **(str. exp.)** |
|---|---|---|---|---|---|
| *Moltbook* | 3 | Logistic | 0.953 | 0.94 | 2.864 |
| | 5 | Logistic | 0.946 | 0.94 | 3.127 |
| | 10 | Logistic | 0.953 | 0.95 | 2.601 |
| *Rogue (engagement)* | 1 | Logistic | 0.996 | 0.986 | 1.637 |
| | 3 | Logistic | 0.996 | 0.986 | 1.709 |
| | 10 | Logistic | 0.996 | 0.985 | 1.421 |
| *LLM-Schelling (informed; entr.)* | 50 | Stretched Exp. | 0.976 | 0.986 | 1.253 |
| | 100 | Stretched Exp. | 0.978 | 0.986 | 1.295 |
| | 200 | Stretched Exp. | 0.975 | 0.983 | 1.224 |
| *LLM-Schelling (uninformed; entr)* | 50 | Logistic | 0.985 | 0.985 | 0.273 |
| | 100 | Logistic | 0.975 | 0.974 | 0.301 |
| | 200 | Logistic | 0.961 | 0.956 | 0.323 |

Table 6: Sensitivity of best-fit results to the minimum-activity threshold $m_{\min}$ (subsampling). For each configuration we report the model preferred by AIC ($\Delta\text{AIC} > 2$), the goodness of fit $R^2$ for both reference forms, and the stretched-exponential stretching exponent $\beta$. Logistic remains the preferred form across all configurations; $R^2$ values remain within a range of 0.01-0.02 for each system. The stretched-exponential $\beta$ confirm the fit: Moltbook's $\beta \gg 1$ indicates a compressed exponential, while Rogue's $\beta$ hovering around 1.5 signals collapse to a pure-exponential form.

Remark from Table 6 that the logistic form is the preferred reference form for most informed systems is robust across all tested $m_{\min}$ values. This confirms that our characterisation of Moltbook and Rogue as phase-transition-like is not a product of our choice of sampling strategy. Second, the diagnostic use of $\beta$ discussed in Section 4 is stable: Moltbook's $\beta$ sits near the fitter's upper bound in every configuration, while Rogue's $\beta$ pins to 1 in every configuration. These are two signatures of shape mismatch which also persist across choice of $m_{\min}$.

## I NULL BASELINES: FORMAL DETAILS

Section 4 of the main paper introduces our marginal-null baseline; here we give the formal details.

**Setup.** Fix a window ending at time $t$. Let $A_t$ be the set of active agents in the window, and $C$ the set of categories (submolts, neighbour-types, or post IDs, depending on the system). The observed data is a matrix $n_{i,j}$ giving agent $i$'s count in category $j$. The per-agent marginal is $n_{i,\cdot} = \sum_j n_{i,j}$; the per-category marginal is $n_{\cdot,j} = \sum_i n_{i,j}$; the total is $N = \sum_{i,j} n_{i,j}$.

**Analytic null.** The analytic estimate of the null mean order parameter is the Shannon entropy of the category-marginal distribution:

$$\Phi_{\text{null}}^{\text{analytic}} = -\sum_{j \in C} \frac{n_{\cdot,j}}{N} \log_2 \frac{n_{\cdot,j}}{N}. \tag{4}$$

This is the entropy of the overall category-popularity distribution. It corresponds to the expected value of the mean-agent-entropy order parameter, if each agent's activity were drawn from the shared category marginals independently of agent identity. That is, as if the agents had distributed

their activity proportional to the overall category popularity, and without any per-agent preferences. Equation 4 is closed-form and cheap to compute.

**Sampled null.** The analytic mean above ignores finite-population fluctuations, particularly relevant when many agents have small counts. To correct for this, we sample synthetic joint distributions $n_{i,j}^{(s)}$ that preserve the observed marginals $n_{i,\cdot}$ and $n_{\cdot,j}$ in expectation and compute the mean-agent-entropy order parameter for each. Concretely, for each category $j$, we distribute $n_{\cdot,j}$ interactions across agents in proportion to the per-agent marginals $n_{i,\cdot}$ (sampling without replacement), and repeat for each category. This is a proportional-allocation approximation of the bipartite configuration model, chosen for its tractability; it recovers the same expected order parameter as the analytic null while introducing the finite-population variance that the analytic estimate omits. The $95\%$ null band reported in Section 5 is the $\{2.5, 97.5\}$ percentile range over $S = 100$ draws.

**Interpreting the null.** The null does not represent a plausible 'unsegregated' system in any behavioural sense. It is a statistical measure answering the following: 'given the observed marginal counts, what value of the order parameter $\Phi$ would we expect if the joint agent-category distribution were maximally mixed while preserving the marginals?' Observed values well outside the null band indicate that the observed joint distribution is not attributable to marginal popularity alone. In other words, that agents have per-agent preferences distinguishable from population-level popularity/frequency effects. This is the strict definition of what we refer to when we indicate 'statistically significant self-organisation'.

**Bad-reach null.** Rogue's bad-reach metric is a scalar ratio, not an entropy, and does not admit the same marginal-null construction. Instead we use a direct *ambient rate*: the fraction of *all* interactions across the run that target malicious content:

$$\Phi_{\text{null}}^{\text{bad reach}} = \frac{\sum_{i,j:j\in M} n_{i,j}}{\sum_{i,j} n_{i,j}}, \tag{5}$$

where $M \subseteq C$ is the set of malicious posts. This measures the ambient rate at which malicious content captures network attention in the aggregate, and across both benign and malicious agents. Benign observed bad reach below this ambient rate indicates safety-tuning resistance (Section 5.3). This null does not have an associated band because it is computed from a single fixed set of totals.

## J Extended Ablation Results II: frozen system dynamics

This section provides the full results described in Section 5.5. We noted that GovSim and ChatEval do not exhibit measurable population-level dynamics under the conditions tested, since the various $\Phi$ we measured had little-to-no-change across ticks (Table 7). Namely, both systems' $\Phi$ remain at fixpoints from the outset: in GovSim, agents solve the harvest coordination in a single round; and in ChatEval referees decide by round 0 and deliberation does not alter their votes. This, from our framework, indicates that no *actual* system dynamics (cooperation or consensus via deliberation) are induced in the system, since no trajectories are found. We argue that this is because the systems are essentially a task with an easily-found objective (i.e., a consensus on harvest, or a response's label) which the agents solve immediately. Systems that self-organise without such a target (LLM-Schelling, Moltbook, Rogue) produce measurable trajectories; systems with an easily locatable target do not. These conclusions were corroborated by independent tests on the results of the configuration sweeps, which we describe next.

### J.1 GovSim: full sweep

Our sweeps for GovSim are as described in Appendix D. The *default* setup totalled 27 runs (3 seeds $\times$ $N$ $\times$ $L$) and is equivalent to the original work's. *Hard* totalled 36 runs, due to the inclusion of $N = 5$. The mechanics and parameters were chosen to increase coordination difficulty and force (or attempt to force) dynamics not observed under the much easier *default* setup. For each experiment run we computed four order parameters per round: *norm entropy* (Shannon entropy of the distribution over agents' stated proposals), *norm variance* (variance of proposals), *coalition entropy* (entropy of actual harvest values), and *coalition variance* (variance of harvests). All order parameters were normalised to $[0, 1]$.

| System | Order parameter | $n_{\text{runs}}$ | $\Phi_0$ | $\Phi_\infty$ | Result |
|---|---|---|---|---|---|
| *GovSim (default)* | Norm entropy | 27 | 0.04 | 0.04 | Frozen: single-round convergence |
| *GovSim (harder)* | Norm entropy | 36 | 0.06 | 0.05 | Frozen: no dynamics on metrics |
| *GovSim (harder)* | Coalition variance | 36 | 0.31 | 0.02 | Partial: N=5 shows $\beta \sim$ 1.8-3.0 |
| *ChatEval (all)* | Verdict entropy | 81 | 0.07 | 0.07 | Frozen: consensus at onset |
| *ChatEval (disagreement)* | Verdict entropy | 21 | 0.81 | 0.81 | Frozen: persistent 3-1 or 2-2 splits |

Table 7: Order parameters in the two 'frozen' (trajectory-less) systems. $\Phi_0$ and $\Phi_\infty$ are median values across runs at round 0 and run end respectively. In all reported cells except the GovSim coalition-variance restriction to $N = 5$, observed and endpoint values are indistinguishable, indicating that the system is at its dynamical fixpoint at the first observation and remains there.

***Default*.** In *default*, trajectories collapsed to zero within a single round for both norm and coalition metrics. Across all 27 runs, 79% of trajectories failed to fit any reference form, with the fitter pinning to boundary $\beta$ values ($\beta \to 0.05$ or $\beta \to 5.0$) in the remaining 21%. These are trivial trajectories: agents converge on a uniform harvest amount by round 1 and hold it for the remaining 11 rounds. Median $\Phi_0 = 0.04$ (norm entropy) and median $\Phi_\infty = 0.04$; the difference is zero within numerical noise.

***Hard*.** Under *hard*, where we induced stock collapse as an eventual outcome, trajectories remain effectively flat. Table 8 summarises the per-cell $\beta$ distribution across this sweep. Of 144 fitted trajectories, 79.2% are dropped (flat trajectory, poor fit, or trajectory too short), and among those that fit, $\beta$ pins to fitter boundaries in the majority of cells. Remark that *hard* not producing dynamics implies that the first possible explanation (that *default* is too easy and harder economics induce relaxation) can be rejected. In both setups the systems remained within a fixpoint, indicating that **no system evolution takes place**.

| Metric | Cells fit | Median $\beta$ | IQR $\beta$ | $\beta$ At boundary | Dynamics obs. |
|---|---|---|---|---|---|
| *Norm entropy* | 9/36 | 0.17 | [0.05, 5.00] | 8/9 | 1/9 (L1, N=10) |
| *Norm variance* | 9/36 | 0.16 | [0.05, 5.00] | 7/9 | 2/9 |
| *Coalition entropy* | 5/36 | 5.00 | [4.71, 5.00] | 5/5 | 0/5 |
| *Coalition variance* | 7/36 | 1.85 | [0.05, 5.00] | 4/7 | 3/7 (all N=5, L1) |

Table 8: GovSim for the *hard* setup: fit distribution across 36 runs per metric. 'Cells fit' is the number passing the minimum-quality threshold ($R^2 \geq 0.5$ and $\beta$ not at boundary). Remark that this criterion is substantially relaxed and in the main paper we use $> 0.75$. Dynamics observed indicate fits with $\beta$ in the non-boundary range $[0.5, 3.0]$. Only $N = 5$ under $L_1$ on coalition variance produces multiple non-boundary fits, albeit $R^2$ remains below the main paper threshold.

### J.2 ChatEval: full sweep

In ChatEval we ran a sweep over $p_{\text{info}} \times L \times 3$ seeds, for a total of 58 runs. The full benchmark contains 68 items; each run processes all items. This yields $81 \times 68 = 5,508$ possible trajectories ('item-outcomes') in total.

**Per-item convergence.** Across all item-outcomes, verdict entropy at round 0 (before deliberation) and at round 2 (after deliberation) differ by less than 0.02 bits (median 0, mean $-0.001$). Referee accuracy remains essentially unchanged: median $0.750$ at round 0 and median $0.750$ at round 2. Deliberation moves neither entropy nor accuracy; consensus is reached before deliberation begins.

**Structure of dissent.** Filtering to items with non-unanimous round-0 verdicts identifies 21 unique items where the referees ever disagree. Table 9 shows the modal disagreement pattern for each of the 21 non-unanimous items across all runs. The pattern is consistent: 20 out of the 21 items produce a persistent 3-1 split, and the last item produces a deadlock. We did not observe item transitions between consensus and disagreement across the item-outcomes tested. Indeed, no item's disagreement pattern shifts (e.g., from 3-1 to 2-2) except in one instance, where the dominant pattern is stable at 50% of runs. Deliberation and information dilution move none of these items. Thus, the fixpoint yielded by $\Phi$ is verifiable.

Aside, in the 3-1 pattern, the same referee (persistently across runs)[8] dissents from the other three. Which referee it is depends on the item–sometimes the *scientist*, sometimes the *critic*–but for a given item the *identity* of the dissenter is stable across seeds. This suggests that ChatEval's dissent is a property of each referee's prior alignment with the underlying task, **not a stochastic outcome of the deliberation process**.

| Item ID | Dominant split | Runs | Stability |
|---|---|---|---|
| mtbench__q103__turn2 | 2-2 | 72 | 74% |
| mtbench__q144__turn1 | 3-1 | 59 | 73% |
| mtbench__q141__turn1 | 3-1 | 56 | 59% |
| mtbench__q114__turn2 | 3-1 | 32 | 81% |
| rewardbench__hep-rust | 3-1 | 31 | 94% |
| rewardbench__llmbar-GPTInst | 3-1 | 29 | 90% |
| mtbench__q113__turn2 | 3-1 | 17 | 100% |
| mtbench__q145__turn1 | 3-1 | 16 | 50% |
| mtbench__q157__turn2 | 3-1 | 12 | 92% |
| rewardbench__llmbar-neighbor | 3-1 | 10 | 100% |
| rewardbench__math-prm | 3-1 | 6 | 100% |
| mtbench__q115__turn1 | 3-1 | 5 | 100% |
| rewardbench__alpaca-hard | 3-1 | 4 | 100% |
| mtbench__q109__turn1 | 3-1 | 2 | 100% |
| rewardbench__hep-python | 2-2 | 2 | 50% |
| rewardbench__alpaca-easy | 3-1 | 2 | 100% |
| mtbench__q121__turn1 | 3-1 | 1 | 100% |
| mtbench__q116__turn2 | 3-1 | 1 | 100% |
| mtbench__q124__turn2 | 3-1 | 1 | 100% |
| mtbench__q95__turn2 | 3-1 | 1 | 100% |
| mtbench__q92__turn2 | 3-1 | 1 | 100% |

Table 9: Non-unanimous items in the ChatEval sweep. *Dominant split* is the modal referee vote distribution at round 2 across all runs where the item appeared. An item may not appear in every run since some items produce unanimous consensus and are excluded from this analysis). *Runs* is the number of runs in which the item appeared as non-unanimous; *Stability* is the fraction of those runs where the dominant split was observed. 20 items exhibit a persistent 3-1 pattern, andonly q103__turn2 exhibits a persistent 2-2 deadlock. Deliberation and information dilution do not transition items between consensus and dissent, nor between 3-1 and 2-2 patterns, thus not yielding any type of trajectory in the system.

## K MOLTBOOK POPULARITY TRENDS

In Figure 6 we show a screengrab of Google search trends worldwide for the keyword 'moltbook'. There is a peak around mid-February, but then search interest wanes quickly and stabilises.

[8]The 'same referee' does not mean the same LLM-prompt-temperature combination. Temperatures are sampled i.i.d. ahead of any run, and the prompt changes across trajectories.

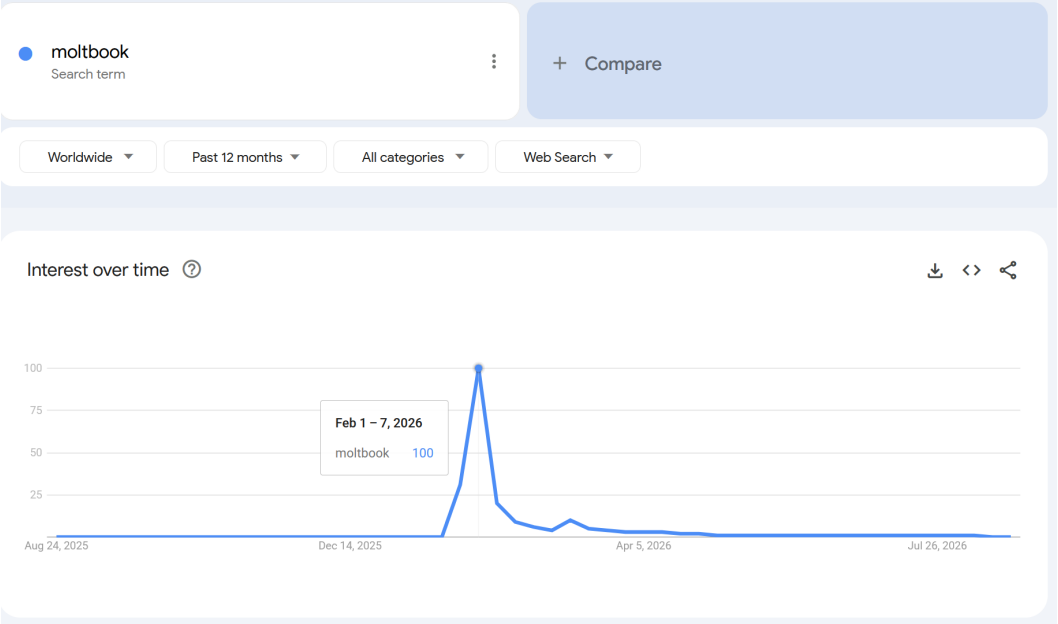



Figure 6: Screenshot of the search interest trends in Google for the word 'moltbook', since its inception and until now. Taken 29 August 2026.