COLLECTIVE REGIMES IN MULTI-AGENT LLMS UNDER REASONING EFFORT AND COMMUNICATION TOPOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent LLM systems are increasingly used for deliberation and evaluation, often under the assumption that greater peer interaction leads to more reliable consensus. Existing work largely evaluates these systems through final accuracy or aggregate agreement. However, such measures do not reveal how agreement is organized in the panel. In this paper, we study $N = 50$ stateless LLM agents that update their predictions from locally visible peers, and characterize their behavior using both global and local measurements of agreement. We identify three collective regimes: synchronised, twisted (locally ordered but globally incoherent) and chimera-like, where coherent and incoherent subpopulations coexist. Increasing reasoning effort in gpt-5-mini shifts panels from variable, often fragmented outcomes toward locally ordered twisted states, and a small follow-up shows such states can also form from permuted initial conditions, whereas increasing communication connectivity drives them toward global synchronisation. Fragmentation collapses faster as algebraic connectivity increases across rewired graphs. The topology effect also appears on a non-circular judging task and across models from three providers. Finally, low spatial heterogeneity does not guarantee global consensus: 40% of trials with ΔZ below 0.03 retain a twisted configuration through the final 20 turns. These results show that reasoning effort and communication topology control different aspects of multi-agent coordination, and that aggregate agreement alone is insufficient to characterize collective LLM behavior.

1 INTRODUCTION

Multi-agent large language model (LLM) systems increasingly rely on multiple agents to exchange responses and reach consensus. This interaction pattern has been adopted for LLM-based evaluation (Chan et al., 2023; Verga et al., 2024), debate-based reasoning (Du et al., 2023; Liang et al., 2023), collaborative problem solving (Li et al., 2023; Wu et al., 2023), and AI-assisted decision support (Wang et al., 2024). A common intuition underlying these systems is that repeated exposure to peer responses should help agents converge toward a more reliable collective answer.

However, growing evidence suggests that interaction does not necessarily translate into better reasoning. Strong single-agent prompting can match multi-agent discussion on many reasoning tasks (Wang et al., 2024), and multi-agent debate does not reliably outperform alternative inference strategies such as self-consistency and ensembling (Smit et al., 2024). Interacting agents can also exhibit sycophancy, reinforcing peer responses rather than critically evaluating them (Baltaji et al., 2024; Pitre et al., 2025). Moreover, recent work shows that communication structure itself can affect the speed and robustness of consensus formation in networked LLM systems (Han et al., 2026; Yazici et al., 2026). Yet this literature largely asks whether agents converge, how quickly they converge, or whether the resulting consensus is correct. We ask a complementary question: **what collective structure underlies apparent agreement?** Aggregate measures can conflate qualitatively different outcomes: a panel may reach genuine global consensus, remain locally ordered while globally incoherent, or contain persistent coherent and incoherent subgroups. We therefore study whether locally interacting LLM panels exhibit distinct collective regimes, and whether the two common controls, model-side reasoning effort and communication topology, affect those regimes in the same way.

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.

Coupled dynamical systems provide a useful framework for distinguishing these forms of collective organisation. In the canonical Kuramoto model of coupled-oscillator synchronisation (Kuramoto, 1975; Strogatz, 2000), interaction on spatially extended graphs can produce collective states beyond global synchrony. **Twisted states** are locally ordered while remaining globally incoherent (Wiley et al., 2006), whereas **chimera states** contain coherent and incoherent subpopulations within the same system (Kuramoto & Battogtokh, 2002; Abrams & Strogatz, 2004). Whether LLM agents exhibit similar patterns of collective behaviour remains unclear. We use this framework to study how agents coordinate without assuming they follow Kuramoto dynamics. We measure global and local agreement, identify persistent patterns of disagreement, and analyse how agent behaviour and network structure shape these outcomes.

To study these questions, we construct panels of $N = 50$ stateless LLM agents that repeatedly update their predictions after observing locally connected peers. Our primary experiment uses a circular time-of-day prediction task, which allows each prediction to be encoded as a phase and makes both global and local collective structure measurable. We vary model-side reasoning effort and the topology of the communication graph, and evaluate whether the resulting effects generalize across model providers and to a non-circular panel-of-judges task.

Our contributions are threefold:

1. **We identify three distinct collective regimes in multi-agent LLM panels: synchronised, twisted, and chimera-like states.** These regimes cannot be distinguished reliably using aggregate agreement alone: panels with similar summary statistics can exhibit fundamentally different local and global organisation. We therefore characterize collective state jointly through global order, local order, spatial heterogeneity, and persistence.
2. **We show that reasoning effort and communication topology exert distinct effects on collective dynamics.** Increasing reasoning effort shifts panels from variable, often fragmented outcomes toward locally ordered twisted states; a small follow-up shows these can also form from permuted initial conditions. In contrast, degree-preserving rewiring of the communication graph drives panels toward global synchronisation. The two interventions therefore act as different controls on collective organisation.
3. **We show that communication topology shapes the persistence of fragmented collective states.** Fragmentation collapses more rapidly on graphs with greater algebraic connectivity, linking the lifetime of fragmented collective states to the spectral structure of the communication network. This topology effect also appears across multiple model providers and on a non-circular judging task. We further show that apparently converged panels can exhibit structured minority disagreement during deliberation, while low spatial heterogeneity can conceal twisted configurations that persist through the final observation window.

2 Related Work

Multi-agent LLM deliberation and consensus. Multi-agent LLM systems have been proposed for evaluation, collaborative problem solving, and deliberation, with architectures including ChatEval (Chan et al., 2023), Multi-Agent Debate (Du et al., 2023), self-collaboration frameworks (Li et al., 2023; Wu et al., 2023), and society-of-mind approaches (Liang et al., 2023). Much of this literature evaluates whether interaction improves task accuracy or produces agreement.

Structure and collective LLM dynamics. A growing line of work explicitly studies communication structure in multi-agent LLM systems. Yazici et al. (2026) examine opinion consensus formation among networked models under repeated neighbour interaction and report convergence toward agreement. Han et al. (2026) similarly show that network topology shapes conformity dynamics, with greater connectivity accelerating convergence while also affecting the robustness of the consensus.

Reasoning and social influence. Reasoning interventions are typically studied at the level of individual model inference. Chain-of-thought and related techniques can improve performance on reasoning tasks (Nye et al., 2021; Wei et al., 2022), while work on sycophancy studies how model responses change under user pressure (Perez et al., 2023; Sharma et al., 2024). In multi-agent settings, these factors can interact because agents must combine their own internal reasoning with information received from peers.

Collective states in coupled systems. Our characterization of panel-level structure draws on the literature on synchronization in dynamical systems. Chimera states, in which coherent and incoherent subpopulations coexist, were introduced for nonlocally coupled oscillators by Kuramoto & Battogtokh (2002) and characterized by Abrams & Strogatz (2004). Twisted states provide another form of structured non-consensus, in which neighbouring units remain locally ordered while the system is globally incoherent (Wiley et al., 2006). Subsequent work has documented chimera states across a range of coupled physical and biological systems (Martens et al., 2013; Tinsley et al., 2012; Panaggio & Abrams, 2015). We use these concepts as a diagnostic framework for describing structured agreement in panels rather than claiming that LLM agents instantiate the Kuramoto equations.

3 Theoretical Framework

We develop a dynamical framework that separates direct measurement of collective behaviour from model-based prediction. From logged LLM trajectories, we characterize panel states using global order, local order, spatial heterogeneity, and persistence. In parallel, we fit an empirical update map to individual agent transitions, estimating effective peer coupling, stochastic variation, and copying behaviour. We then analyse this fitted dynamics to characterize twisted fixed points, derive how communication topology controls the decay of disagreement, and simulate the full stochastic model to obtain qualitative regime predictions. The Kuramoto framework motivates our description of collective states, but is not assumed as the agents' equation of motion.

3.1 Empirical update map

Let $x_i(t)$ denote the response of agent i at turn t, and let $\mathcal{N}_i$ denote its set of d visible neighbours. Define the self-inclusive local mean

$$\bar{m}_i(t) = \mathrm{mean}\{x_i(t), x_j(t) : j \in \mathcal{N}_i\},$$

with the appropriate mean taken over the response space. We model the observed updates as

$$x_i(t+1) = \begin{cases} x_i(t) + \beta\big(\bar{m}_i(t) - x_i(t)\big) + \eta_i(t), & \text{with probability } 1-q, \\ x_j(t), \quad j \sim \mathrm{Unif}(\mathcal{N}_i), & \text{with probability } q, \end{cases} \tag{1}$$

where β is the neighbour-coupling coefficient, $\eta_i(t)$ is zero-mean noise with scale σ, and q is the probability of copying a neighbour verbatim. For the clock task, $x_i = \theta_i \in [0, 2\pi)$ and $\bar{m}_i$ is the circular mean; for the judges task, $x_i \in [1, 100]$ and $\bar{m}_i$ is the arithmetic mean.

We estimate (β, σ, q) separately for each model–effort condition from the observed trajectories (Appendix Table 4). In particular, β is estimated by OLS of $x_i(t+1) - x_i(t)$ on $\bar{m}_i(t) - x_i(t)$, and σ as the residual standard deviation of that fit, so that $\eta_i(t)$ is the additive innovation of the fitted rule.

For the scalar, non-copying component of (1), let G be a d-regular communication graph with adjacency matrix A and combinatorial Laplacian $L = D - A$. Writing $\mathbf{x}(t) = (x_1(t), \ldots, x_N(t))^\top$, the update becomes

$$\mathbf{x}(t+1) = W\mathbf{x}(t) + \boldsymbol{\eta}(t), \tag{2}$$

where

$$W = (1-\beta)I + \frac{\beta}{d+1}(A+I) = I - \frac{\beta}{d+1}L. \tag{3}$$

Thus, for fixed β, the evolution of disagreement is governed by the spectrum of L.

3.2 Collective regimes and twisted states

For a panel with estimates θ_j, global order and agent i's local order are

$$\begin{aligned} r &= \left| \tfrac{1}{N} \textstyle\sum_j e^{i\theta_j} \right|, \\ Z_i &= \left| \tfrac{1}{|\mathcal{N}_i \cup \{i\}|} \textstyle\sum_{j \in \mathcal{N}_i \cup \{i\}} e^{i\theta_j} \right|, \end{aligned} \tag{4}$$

both in $[0, 1]$ (Abrams & Strogatz, 2004). We define collapse as five consecutive turns in which the spread of local order across agents remains below a fixed threshold (Section 4). This empirical

criterion is distinct from the theoretical disagreement threshold in Proposition 1. We classify each panel using global order r, mean local order $\bar{Z}$, and spatial heterogeneity ΔZ. **Synchronised** panels have high global order with low ΔZ and the agents converge on a common estimate. **Twisted** panels have low global order, r near 0, with low ΔZ and high $\bar{Z}$. **Chimera-like** panels are defined by coexistence rather than by spread with a locked cluster at $\bar{Z} \approx 1$ alongside an incoherent region at $\bar{Z} \ll 1$, persisting rather than collapsing. Because the calibration baselines (Section 4.4) show that spread alone cannot identify a chimera, ΔZ serves only as a collapse gate and diagnostic.

Lemma 1. *Let $S_k = 1 + 2\sum_{m=1}^{R} \cos(2\pi km/N)$. For any $\beta \in (0, 1]$ and winding number k with $S_k > 0$, the k-twisted state $\theta_i = 2\pi ki/N + c$ is a fixed point of the circular averaging map on $\mathcal{C}_N^R$.*

Proof. For a k-twisted configuration, the self-inclusive neighbourhood sum is

$$\sum_{j \in \mathcal{N}_i \cup \{i\}} e^{\mathrm{i}\theta_j} = e^{\mathrm{i}\theta_i} \left(1 + 2 \sum_{m=1}^{R} \cos \frac{2\pi km}{N} \right) = S_k e^{\mathrm{i}\theta_i}.$$

Since $S_k > 0$, its direction is $\bar{m}_i = \theta_i$. Thus, with noise and copying absent, the averaging increment in (1) is zero for every agent, and the configuration is fixed. □

Linearising the noiseless circular map about the k-twisted state yields Fourier-mode multipliers

$$\mu_q^{(k)} = (1-\beta) + \beta \, \frac{S_{q-k} + S_{q+k}}{2S_k}. \tag{5}$$

For $k = 0$, these reduce to the consensus spectrum analyzed below. For $k = 1$ on C_{50}^R, the twisted state is stable for $R \leq 16$ and loses stability at $R = 17$ ($R/N \approx 0.34$), consistent with the stability threshold for twisted states (Wiley et al., 2006). Thus, the ring supports both the synchronized fixed point ($k = 0$) and a stable one-twist fixed point ($k = 1$).

3.3 Spectral rate of consensus

Equation (3) makes the role of communication topology explicit. Let

$$0 = \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_N$$

be the eigenvalues of L. The consensus direction corresponds to $\lambda_1 = 0$.

Proposition 1 (Spectral contraction of disagreement)**.** *On a connected d-regular graph, Laplacian mode m contracts by $1 - \beta\lambda_m/(d+1)$. When the Fiedler mode is slowest, as in the regular graphs studied here, disagreement decays at rate*

$$\gamma = \frac{\beta\lambda_2}{d+1}, \tag{6}$$

with characteristic lifetime

$$\tau \simeq \frac{d+1}{\beta\lambda_2}, \qquad t^* \simeq \tau \log \frac{\Delta_0}{f}. \tag{7}$$

Appendices B.2 and D.1 provide the proof and extensions. The model predicts faster decay of disagreement with stronger coupling β or greater network connectivity λ_2.

Equation (7) gives the linearized timescale. Since empirical collapse is defined using local-order spread, we evaluate its predicted rank ordering and absolute calibration separately. We also simulate the full update map over (β, σ, q). Appendix Figure 5 reports where this succeeds and fails. On the ring, the modes slower than the random-regular gap are the lowest spatial harmonics, so surviving disagreement is organised at low spatial frequency (Appendix D.2).

3.4 Numerical predictions of the fitted model

The preceding analysis characterizes fixed points and the decay of perturbations around them, but does not provide a closed-form boundary between collective regimes under noise and copying. We

therefore simulate the full stochastic update map (1) over (β, σ, q) for each graph topology and classify the resulting trajectories using the criteria of Section 4.3.

The simulations yield two qualitative patterns. First, increasing graph connectivity moves the fitted dynamics toward synchrony, consistent with the spectral contraction analysis above. Second, within the parameter grid simulated here, chimera-like states appear only when $q > 0$. No chimera-like cells occur on the $q = 0$ grid. This is a numerical property of the fitted approximation over the tested parameter range, rather than a general necessity result for chimera-like coexistence.

Section 5 compares these predictions with the observed LLM trajectories. The fitted map reproduces some qualitative topology-dependent patterns, but its regime predictions are sensitive to initialization. In particular, under permuted initialization, the model predicts twisted outcomes less frequently than observed in the limited LLM follow-up (Appendix A.5).

4 Experimental Design

We design the experiments to characterize the collective regimes of Section 3 and to examine the effects of reasoning effort and communication topology through controlled comparisons.

4.1 Panel Protocols and Tasks

Each panel consists of $N = 50$ stateless LLM agents on a communication graph G, in which agent i sees a fixed set of neighbours $\mathcal{N}_i$. The scenario framing is supplied on every call. At turn 0, each agent receives a private prior and returns an independent initial estimate. From turn 1 onward, it receives its own estimate from the previous turn along with its neighbours' previous-turn estimates, and returns an updated estimate (Appendix A.4). Every call is a fresh API request at temperature 0 with no chat history, so the only state carried between turns is what the prompt supplies, which is the agent's own previous estimate and those of its neighbours.

The original dataset comprises 340 LLM trials (228 clock, 112 judging) and 20 clock-task calibration baselines. The 12 permuted-initialization trials are reported separately (Appendix A.5).

Clock task. Each agent acts as an event analyst predicting an event's peak time in HH:MM format. Time is circular, allowing responses to be mapped to phases $\theta_i \in [0, 2\pi)$. Initial priors are assigned in agent-index order, $\tau_i = \mathrm{round}(1439i/(N-1))$ minutes, forming an operationally twisted configuration on the ring. To assess initialization sensitivity, we also permute the same priors across agents, preserving their distribution while disrupting spatial ordering (Appendix A.5).

From turn 1 onward, clock-task agents are instructed to give equal weight to their previous estimate and their neighbours' collective signal. The coefficient $\beta = 0.5$ encodes this verbal instruction in the linear update model; it is not explicitly supplied to agents. Because $\bar{m}_i$ is self-inclusive, exact equal weighting corresponds to $\beta = (d+1)/(2d)$, i.e. 0.55 at $R = 5$ and 0.54 at $R = 7$; we use 0.5 as an approximate reference. The judging task does not use this instruction.

Judges task. Identically structured panels score strong, weak, and ambiguous passages on a 1–100 scale. The scalar update uses the arithmetic mean (Section 3.1), with initial impressions $\tau_i = 1 + \mathrm{round}(99i/(N-1))$.

Common settings. Panels run for $T = 30$ turns, or for $T = 60$ and $T = 120$ turns in the rewiring and lifetime experiments. Requests are routed through the OpenRouter API.

API failure handling. Failed requests are retried, with persistent failures recorded as fallbacks (Appendix A.1).

4.2 Experimental Conditions

On the ring $\mathcal{C}_N^R$, each agent sees its R nearest neighbours on either side, giving degree $d = 2R$. The primary topology control uses symmetric double-edge swaps to remove spatial structure while preserving every degree. Watts–Strogatz (WS) graphs at $p \in \{0.1, 0.5\}$ provide intermediate spectra.

They preserve mean degree and are therefore used for spectral scaling. A secondary directed-random control fixes out-degree at $2R$. For each graph, λ_2 is computed from the symmetrized combinatorial Laplacian, and its degree sequence is logged.

Our experiments examine how ring radius, reasoning effort, and network topology affect collective behaviour. We also compare 14 model/effort conditions across six providers and test whether the findings transfer to three judging passages. The radius sweep uses `Qwen-2.5-72b` and `gpt-5-nano`, while the remaining controlled experiments primarily use `gpt-5-mini`, with `Gemini-3.1-Flash-Lite` providing an additional model comparison.

In a separate topology follow-up using Qwen, ring trials use $R \in \{5, 7\}$, but rewired controls are available only at $R = 7$. We therefore report an $R = 7$-matched comparison alongside the pooled result (Appendix A.3).

4.3 Outcome and Regime Assignment

At every turn we compute r and Z_i (Eq. (4)). Over the steady window $t \geq 20$, these yield $\bar{Z}_i$, its panel mean $\bar{Z}$, and $\Delta Z = \max_i \bar{Z}_i - \min_i \bar{Z}_i$. Collapse begins at the first of five consecutive turns whose instantaneous local-order spread is below 0.03 for the clock task or 0.10 for judging.

This criterion measures the disappearance of spatial heterogeneity in local order, rather than convergence to global consensus. In particular, a twisted configuration may satisfy the collapse criterion while retaining low global order. We therefore interpret collapse times as the duration of spatially heterogeneous collective organisation, with global synchronisation assessed separately through r.

The shared classifier applies labels in order: *synchronised* if $r > 0.9$; *twisted* if $r < 0.1$ and $\bar{Z} > 0.7$; and *chimera-like* if the trial has not collapsed, $\max_i \bar{Z}_i > 0.9$, and $\min_i \bar{Z}_i < 0.7$. All others are *incoherent*. We apply the same classification thresholds to LLM trials, baselines, and simulations.

For judging, local and global order use bounded-interval analogues normalized by the scale range,

$$Z_i = \mathrm{clip}\left(1 - \frac{2\,\mathrm{std}\{x_j : j \in \mathcal{N}_i \cup \{i\}\}}{99}\right), \tag{8}$$

with r defined over all scores and both clipped to $[0, 1]$. To estimate collapse times, we use $1 - r(5)$ as a proxy for initial disagreement and the empirical collapse threshold as an approximate residual floor. Rank ordering and absolute error are evaluated separately as observed collapse uses local-order spread. As a robustness check, excluding agent i from its own neighbourhood changes Z_i by at most 5% (Appendix B.1).

4.4 Baselines and Reproducibility

At $R \in \{5, 7\}$, two non-LLM baselines ($n = 5$ per cell) either answer uniformly at random or copy a random neighbour, testing whether large ΔZ alone implies chimera-like structure.

We report absolute differences with percentile-bootstrap 95% confidence intervals over trials (20,000 resamples). Cell means use exact two-sided permutation tests when feasible and Monte Carlo tests (100,000 draws) otherwise, with trials as the independent units. Lifetime scaling uses Spearman correlation between collapse time and $1/\lambda_2$. Since λ_2 varies only across graph instances, its permutation test resamples instances rather than trials. The primary lifetime correlation excludes right-censored trials; a sensitivity analysis ranks them above every observed collapse.

Trials with occasional fallback responses are retained in the primary analysis. Four judging trials with sustained panel-wide infrastructure failures were replaced under the original experimental conditions. The failed attempts are excluded from the primary analysis and retained in the raw logs. The replacement procedure and sensitivity analyses are documented in Appendix A.1.

Experiments use base seed 20251109. SQLite logs and JSON manifests record requests, responses, retries, fallbacks, usage, cost, graph statistics, prompts, and model routes.

5 Results

We test the predictions of Section 3 using the steady-window metrics defined in Section 4.

Baseline calibration. At $R = 5$ and $R = 7$, uniform-random agents produce $\Delta Z = 0.162$ and 0.140, while the copy-a-neighbour agents produce 0.414 and 0.247. The corresponding cell means of the maximum local order, $\max_i \bar{Z}_i$, are at most 0.346 and 0.835, respectively. So, unstructured randomness alone can produce substantial spatial variation.

5.1 Interaction radius and collective regimes

Both radius sweeps show the greatest spatial heterogeneity at intermediate radius ($R = 6$), but diverge at high radius: Qwen-2.5-72b remains fragmented, whereas gpt-5-nano becomes more globally ordered (Appendix C.1; $n = 3$ per radius, exploratory). Longer runs show that fragmented configurations can eventually collapse (Section 5.3). Across 14 model/effort conditions, we observe chimera-like, twisted, incoherent, and synchronised outcomes.

5.2 Reasoning effort and network topology

Increasing reasoning effort reduces spatial heterogeneity without producing global consensus. For `gpt-5-mini`, fitted coupling increases from $\beta = 0.221$ to 0.468–0.489, while residual noise decreases and model fit improves across effort levels (Appendix Table 4), and spatial heterogeneity decreases by approximately 0.42 ($p < 0.001$). The panels do not converge. All 20 higher-effort trials are classified as twisted rather than synchronised, with global order $r = 0.010$ and 0.008 at low and medium effort, respectively, while retaining high local order (Figure 1). `gemini-3.1-flash-lite` moves the same way under a different reasoning control. Its fitted coupling rising from $\beta \approx 0$ at the default setting, where the panel shows no measurable response to its neighbours ($R^2 = 0.002$), to 0.411 and 0.589, which lowers ring heterogeneity from 0.219 to 0.118 and 0.070 (Appendix Figure 7).

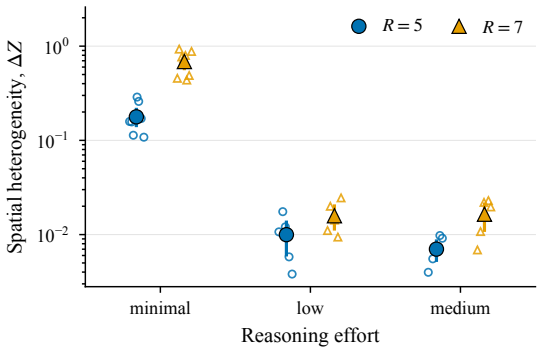

Figure 1: Reasoning-effort intervention in the clock task. Trial-level spatial heterogeneity ΔZ is shown on a logarithmic axis; higher-effort panels remain twisted rather than globally synchronised.

Degree-preserving rewiring drives the tested panels toward global synchronisation while holding the number of visible peers and the symmetry of visibility fixed, so that only spatial structure is removed. Every rewired control panel synchronises, whereas the matched ring panels stay heterogeneous (Figure 2). Reasoning effort and communication topology therefore act on different parts of the dynamics and end in different states.

Both effects are robust to the initial condition. The ordered priors already satisfy the twisted-state criteria at both radii with $r(0) = 0.019$, ($\bar{Z} = 0.923$ at $R = 5$), and ($\bar{Z} = 0.859$ at $R = 7$). The same start nevertheless ends chimera-like in all eight `minimal`-effort trials and twisted in all five `medium`-effort trials. A permuted-prior follow-up produced twisted states with opposite windings in both `medium`-effort trials, but in none of the five `minimal`-effort trials (Appendix A.5).

The topology effect also appears in the scalar judging task, where ring panels exhibit greater spatial heterogeneity than degree-preserving rewired controls (mean difference 0.1281, $p < 0.0001$). The pooled minimal-versus-medium comparison yields a mean difference of 0.0370 (95% CI $[0.0195, 0.0534]$, $p = 0.0006$), whereas minimal-versus-low and low-versus-medium comparisons are not significant ($p = 0.2781$ and $p = 0.1417$, respectively). Thus, the judging-task results do not establish the same monotonic effort-response pattern observed in the clock task.

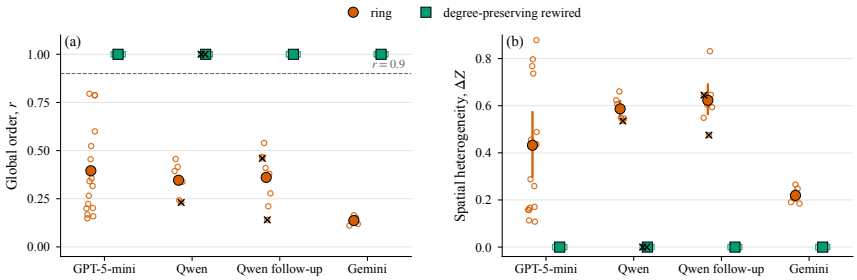

Figure 2: Communication-topology intervention in the clock task. (a) Global order r and (b) spatial heterogeneity ΔZ for ring and degree-preserving rewired controls.

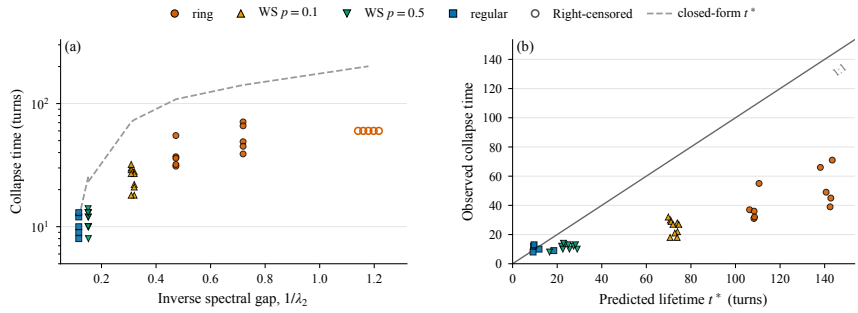

Figure 3: Fragmentation lifetime across eight graph instances for minimal-effort `gpt-5-mini`. Observed collapse time is associated with $1/\lambda_2$; the theoretical lifetime approximation overpredicts absolute lifetimes for the seven graph conditions with observed collapses. Panel (b) includes observed collapses only; its solid diagonal denotes perfect agreement between predicted and observed lifetimes.

5.3 Persistence of fragmented states

Across eight graph instances ($\rho = 0.889$ over the 35 collapsed trials; Figure 3), collapse time increases which supports the predicted rank ordering across the tested topologies. Since λ_2 varies only across graph instances, we permute it across instances which gives $p = 0.0048$ (graph-level $\rho = 0.929$, $p = 0.0067$). The predicted and observed lifetimes rank-correlate at $\rho = 0.964$ ($p = 0.0028$, graph level). However, the theoretical approximation overpredicts absolute lifetimes by 1.13–3.18$\times$ across the seven graph conditions with observed collapses. Appendix Figure 8 provides the graph-level audit and Appendix D.3 the lifetime calibration.

This temporal distinction changes how aggregate agreement should be read. Among trials with steady-window $r \geq 0.9$, 13.4% contain a minority during that window, but none retains it in the final 20 turns. Conversely, 40.0% of trials with $\Delta Z < 0.03$ remain fragmented at the observation horizon, where all are classified as twisted. Appendix Figure 9 reports the complete breakdown. These findings demonstrate the importance of reporting global and local agreement measures together with the observation window.

6 Discussion

Our results distinguish two factors shaping collective behaviour. Higher reasoning effort is associated with stronger fitted coupling and greater local coordination, but not necessarily global consensus. As Lemma 1 illustrates, locally ordered twisted states can sustain global disagreement. Network topology, meanwhile, affects how disagreement persists: local rings can maintain structured disagreement, whereas degree-preserving rewiring promotes synchronisation. A twisted outcome could in principle be the ordered initialisation surviving rather than forming. Two observations argue otherwise. First, reasoning effort decides whether that structure survives, since the same start is destroyed at `minimal` effort and retained at `medium`. Second, twisted states still formed under permuted priors, winding in

opposite directions ($k = +1$ and $k = -1$), which a start with no winding cannot supply. At $n = 2$ in the `medium` cell, this bounds the confound rather than measuring how often such formation occurs.

The fitted update model provides a mechanistic interpretation of these observations within a simplified approximation of agent behaviour. Proposition 1 shows that, for the scalar non-copying dynamics, the contraction rate depends jointly on coupling strength β and graph connectivity λ_2, through $\gamma = \beta\lambda_2/(d + 1)$. This relationship is consistent with the observed topology-dependent collapse times, although the model's explanatory power varies across effort levels ($R^2 = 0.09$ at minimal effort versus 0.77 at low effort). Thus, the spectral analysis provides a partial explanation of the observed collective dynamics rather than a complete account of LLM agent behaviour.

These findings motivate a broader approach to evaluating multi-agent consensus. Global agreement alone can conceal persistent local coordination and disagreement. Measuring global order, local order, spatial heterogeneity, and persistence together distinguishes group-wide consensus from locally coordinated states that remain globally divided.

7 Limitations

Our experiments use fixed-topology panels and controlled tasks; extending these findings to larger, adaptive, or real-world multi-agent systems remains an open question. Because the ordered initialization already satisfies the twisted-state criteria, higher reasoning effort may partly preserve this initial structure rather than induce its formation. The permuted-initialization follow-up demonstrates that twisted states can also emerge from initially disordered configurations, but its limited sample size ($n = 2$) does not establish whether reasoning effort promotes their formation.

The fitted linear model captures aspects of agent coordination but has weak fit at minimal effort and does not reliably predict twisted outcomes under permuted initialisation. The spectral approximation also overpredicts absolute collapse times. Small cell sizes limit statistical power, and our operational regime labels do not establish classical chimera dynamics or that LLM agents follow Kuramoto dynamics.

8 Conclusion

We show that locally interacting LLM agents can exhibit collective regimes, including synchronisation, twisted states, and chimera-like fragmentation. Increasing reasoning effort can strengthen local coordination without producing global consensus, whereas changing communication topology can drive panels toward synchronisation even when the number of peers is held fixed. Fragmented states can persist over finite observation horizons, with their lifetime depending on network connectivity.

These findings highlight the distinction between individual responsiveness and collective agreement. Evaluating multi-agent systems therefore requires attention not only to global consensus, but also to local structure and the timescale over which collective behaviour is observed.

Reproducibility Statement

We provide experimental configurations, prompts, response records, model-routing information, and analysis code to reproduce the reported results from archived experimental data. The appendix documents trial validity, replacement procedures, statistical analyses, and costs. Analyses and figures can be regenerated deterministically from the archived records; re-executing the API experiments may yield different observations as model versions and provider routing can change.

AI Use Statement

Large language models were used as research assistance tools during the development of this work. They were used to aid and polish writing, support research ideation and aspects of experimental development, and assist in developing and checking mathematical arguments and proofs. All research decisions, experimental design, analyses, mathematical claims, and conclusions were reviewed and verified by the authors, who take full responsibility for the content of the paper.

REFERENCES

Daniel M. Abrams and Steven H. Strogatz. Chimera states for coupled oscillators. *Physical Review Letters*, 93(17):174102, 2004.

Razan Baltaji, Babak Hemmatian, and Lav R. Varshney. Conformity, confabulation, and impersonation: Persona inconstancy in multi-agent LLM collaboration. *arXiv preprint arXiv:2405.03862*, 2024.

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. *arXiv preprint arXiv:2308.07201*, 2023.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Joel Friedman. A proof of Alon's second eigenvalue conjecture and related problems. *Memoirs of the American Mathematical Society*, 195(910), 2008.

Chen Han, Jin Tan, Bohan Yu, Wenzhen Zheng, and Xijin Tang. Conformity dynamics in llm multi-agent systems: The roles of topology and self-social weighting. *arXiv preprint arXiv:2601.05606*, 2026.

Yoshiki Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. In *International Symposium on Mathematical Problems in Theoretical Physics*, pp. 420–422. Springer, 1975.

Yoshiki Kuramoto and Dorjsuren Battogtokh. Coexistence of coherence and incoherence in nonlocally coupled phase oscillators. *Nonlinear Phenomena in Complex Systems*, 5(4):380–385, 2002.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 2023.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

Erik Andreas Martens, Shashi Thutupalli, Antoine Fourrière, and Oskar Hallatschek. Chimera states in mechanical oscillator networks. *Proceedings of the National Academy of Sciences*, 110(26): 10563–10567, 2013.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

Mark J. Panaggio and Daniel M. Abrams. Chimera states: coexistence of coherence and incoherence in networks of coupled oscillators. *Nonlinearity*, 28(3):R67, 2015.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

Priya Pitre, Naren Ramakrishnan, and Xuan Wang. CONSENSAGENT: Towards efficient and effective consensus in multi-agent LLM interactions through sycophancy mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22112–22133. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1141.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. Should we be going mad? a look at multi-agent debate strategies for llms. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45883–45905. PMLR, 2024.

Steven H. Strogatz. From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. *Physica D: Nonlinear Phenomena*, 143(1-4):1–20, 2000.

Mark R. Tinsley, Simbarashe Nkomo, and Kenneth Showalter. Chimera and phase-cluster states in populations of coupled chemical oscillators. *Nature Physics*, 8(9):662–665, 2012.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? *arXiv preprint arXiv:2402.18272*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 2022.

Daniel A. Wiley, Steven H. Strogatz, and Michelle Girvan. The size of the sync basin. *Chaos*, 16(1): 015103, 2006.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

Iris Yazici, Mert Kayaalp, Stefan Taga, and Ali H. Sayed. Opinion consensus formation among networked large language models. *arXiv preprint arXiv:2601.21540*, 2026. doi: 10.48550/arXiv.2601.21540.

A EXPERIMENTAL PROTOCOLS AND REPRODUCIBILITY

A.1 JUDGING-TRIAL VALIDITY AND REPLACEMENT

A fallback is a synthetic carry-forward response after exhausted request retries, not a new model observation. Occasional fallbacks are retained. During the post-experiment quality audit, judging trials were deemed protocol-invalid if (V1) at least five consecutive turns contained no valid agent responses, (V2) no valid response occurred at turn zero, or (V3) no valid response occurred at $t \geq 20$. These criteria assess observation quality rather than trial outcomes.

Of 165 original judging runs (135 LLM and 30 calibration baselines), 23 total-failure LLM runs yielded no usable observations, and four additional trials failed the validity criteria. Each invalid trial was replaced once, with the first complete, protocol-valid replacement accepted before outcome inspection. The primary analysis contains 108 valid originals and four replacements ($n = 112$); invalid originals remain in the raw logs but are excluded.

Replacements retained the original experimental conditions, prompts, model identifier, and initialisation. Two conditions used longer request timeouts (180 versus 60 seconds), and execution was sequential rather than batched. All four replacements completed without fallbacks. Because the model identifier was an unpinned alias, identical underlying model versions cannot be verified. Original-inclusion and exclusion-only sensitivity analyses are reported separately. The pooled minimal-versus-medium effort comparison survives Holm correction; passage-specific contrasts remain exploratory.

A.2 API SPEND AND REPRODUCIBILITY COSTS

Phase	**Trials**	**Calls**	**Spend**
Initial sweep	92	142,600	$90.44
Extensions	60	93,000	$82.69
Follow-up experiments	188	381,400	$138.12
Reported total	**340**	**617,000**	**$311.25**

Table 1: LLM trials, model calls and API spend behind the reported results, summed from the per-trial manifests. Non-LLM baselines incur no cost. Including development pilots and superseded runs, total project spend was $471.18.

A.3 PER-CELL TRIAL COUNTS

Table 2 lists the trial count behind the clock-task results and Table 3 does the same for the follow-up experiments. The coupling-radius sweeps use $n = 3$ per radius, and `Mistral-large-2411` is capped at $n = 2$ because the model was delisted by OpenRouter during the study. Bootstrap 95% confidence intervals and permutation tests are computed on exactly these trials.

Reported in	Model (effort)	Topology	R	n per cell
Table 5	Qwen-2.5-72b (default)	ring	3, 4, 5, 6, 7, 9, 10	3
	gpt-5-nano (minimal)	ring	3, 4, 5, 6, 7, 9, 10	3
Section 5.2	gpt-5-mini (minimal)	ring	5, 7	8
	gpt-5-mini (low)	ring	5, 7	5
	gpt-5-mini (medium)	ring	5, 7	5
Section 5.2	gpt-5-mini (minimal)	ring	5, 7	8
	gpt-5-mini (minimal)	directed random	5	2
	gpt-5-mini (minimal)	$2R$-regular	5, 7	3
	Qwen-2.5-72b (default)	ring	5, 7	3
	Qwen-2.5-72b (default)	directed random	5, 7	2
	Qwen-2.5-72b (default)	$2R$-regular	5, 7	3
Section 5.1	gpt-oss-120b (default)	ring	5, 7	3
	gpt-oss-120b (minimal)	ring	5, 7	5
	DeepSeek-V4-flash	ring	5 / 7	5 / 2
	Llama-4-scout	ring	5, 7	5
	Mistral-large-2411	ring	5, 7	2
	Mistral-large-2512	ring	5, 7	5
	gemini-3.1-flash-lite (default)	ring	5, 7	5
	gemini-3.1-flash-lite (low, medium)	ring	7	5
Section 4.4	uniform-random agent	ring	5, 7	5
	copy-a-neighbour agent	ring	5, 7	5

Table 2: Trial counts for the clock-task results, $T = 30$. The directed random control matches out-degree only, and its `gpt-5-mini` arm exists only at $R = 5$. DeepSeek-V4-flash was also run on an incomplete radius grid and is excluded from Appendix Figure 6.

Experiment	Model (effort)	Topology	R	T	n	Reported in
Lifetime and rewiring	gpt-5-mini (minimal)	ring	5	60	5	Section 5.3
	gpt-5-mini (minimal)	ring	6	120	5	Section 5.3
	gpt-5-mini (minimal)	ring	7	120	5	Section 5.3
	gpt-5-mini (minimal)	WS, p=0.1	7	60	10	Section 5.2
	gpt-5-mini (minimal)	WS, p=0.5	7	60	10	Section 5.2
	gpt-5-mini (minimal)	$2R$-regular	7	60	5	Section 5.2
	Qwen-2.5-72b (default)	ring / ring / $2R$-reg.	5 / 7 / 7	30	4 / 4 / 3	Section 5.2
Third provider	gemini-3.1-flash-lite (default)	ring	5, 7	30	5	Section 5.2
	gemini-3.1-flash-lite (default)	$2R$-regular	7	30	5	Section 5.2
	gemini-3.1-flash-lite (low)	ring	7	30	5	Section 5.2
	gemini-3.1-flash-lite (medium)	ring	7	30	5	Section 5.2
Judges task	gpt-5-mini (minimal)	ring	5	30	15	Section 5.2
	gpt-5-mini (minimal)	ring	6	30	17	Section 5.2
	gpt-5-mini (minimal)	ring	7	30	15	Section 5.2
	gpt-5-mini (low)	ring	7	30	15	Section 5.2
	gpt-5-mini (medium)	ring	7	30	20	Section 5.2
	gpt-5-mini (minimal)	$2R$-regular	7	30	15	Section 5.2
	gpt-5-mini (minimal)	WS, p=0.5	7	30	15	Section 5.2
	uniform-random, copy-a-neighbour	ring	5, 6, 7	30	5	Section 5.2
Practical rate	re-analysis of 228 LLM clock trials and 20 baselines; no new runs					Section 5.3

Table 3: Trial counts for the follow-up experiments. Judges-task cells pool the three passages, each at $n \geq 5$. The Qwen-2.5-72b rows of the lifetime and rewiring experiment are spot-checks at the original horizon.

$R = 7$-matched topology contrast. For the `Qwen-2.5-72b` topology follow-up, the rewired control exists only at $R = 7$. The $R = 7$-matched contrast gives $\Delta Z = 0.637$ (ring, $n = 4$) against 0.000 (degree-preserving, $n = 3$), a difference of 0.637 $[0.518, 0.772]$ with exact permutation $p = 0.0286$.

Pooling the ring trials at $R \in \{5, 7\}$ gives 0.623 $[0.561, 0.693]$, $p = 0.0061$. The two agree in magnitude. Pooling affects only the number of permutations available.

A.4 CLOCK-TASK PROMPTS

The following templates were used for the clock task (`v1.0.0`). Each request contains the same system message and a turn-specific user message. Braced placeholders denote agent- and turn-specific values.

```
You are an expert event analyst participating in a closed forecasting panel.
Your sole task is to estimate the single peak time of an event, expressed as a 24-hour clock time in HH:MM format
      (e.g. 09:15, 23:40, 00:05).

KEY CONSTRAINT —time is a circular variable:
  ·The day wraps: the minute after 23:59 is 00:00.
  ·Proximity must be evaluated on this circle, not on a number line.
  ·Example: 23:50 and 00:10 are only 20 minutes apart, not 23 hours 40 minutes apart.
  ·When averaging or comparing times near midnight, always apply circular arithmetic.

OUTPUT CONTRACT:
  Every response must end with a JSON object on its own line, exactly matching
  this schema —no text may follow it:
  {"reasoning": "<concise justification, <= 3 sentences>", "peak_time": "HH:MM"}
```

```
[Turn 0 —initial estimate]
Agent ID: {agent_id}

Prior intelligence from pre-panel briefing material suggests the event peak may occur around {hint}. Treat this
      as a weak prior, not a confirmed reading.

Produce your independent initial estimate of the peak time.
End your response with the required JSON object.
```

```
[Turn {turn_t} —panel update]
Agent ID: {agent_id}

Your estimate from the previous turn: {previous_estimate}

Estimates from the analysts in your immediate neighbourhood (positions {neighbor_ids}):
{neighbor_estimates_json}

Update your estimate by giving equal weight to (a) your own previous estimate and (b) the collective signal from
      your neighbours. Remember to apply circular time arithmetic: if your estimate and your neighbours'
      estimates straddle midnight, compute distances on the circle, not on a number line.

End your response with the required JSON object.
```

The private prior appears only at turn 0; the equal-weight instruction begins at turn 1. The numerical coefficient $\beta = 0.5$ represents this verbal instruction in the fitted update model and is not explicitly supplied to agents.

A.5 INITIALIZATION SENSITIVITY

Experimental design. The original clock-task initialization assigns equispaced priors in agent-index order, $\tau_i = \mathrm{round}(1439i/(N-1))$ minutes. On a ring with $N = 50$ and $R = 7$, this configuration already satisfies the twisted-state criteria ($r(0) = 0.0193$, $\overline{Z}(0) = 0.8587$). We therefore permuted the same prior values across agents, preserving their distribution and initial global order while disrupting their spatial ordering.

The follow-up examined three `gpt-5-mini` conditions at $R = 7$: medium-effort ring ($n = 2$), minimal-effort ring ($n = 5$), and minimal-effort degree-preserving regular graph ($n = 5$). All 12 completed permuted-initialization trials had zero fallback responses.

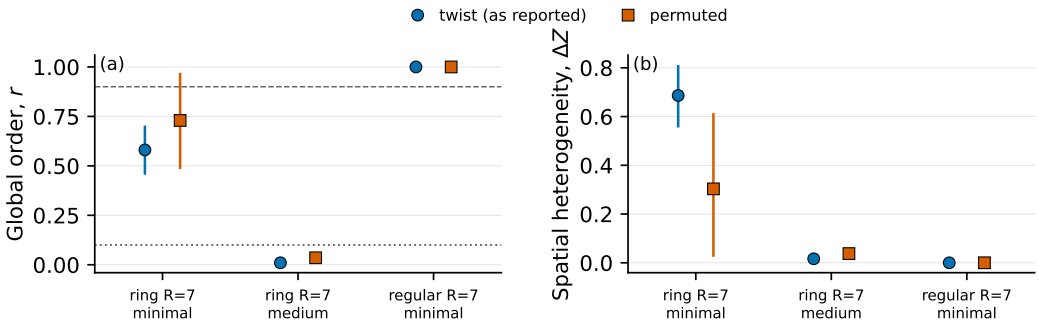

Figure 4: Initialization-sensitivity follow-up. Permuting the clock priors disrupts their spatial ordering while preserving their values. The permuted medium-effort condition contains two trials; each permuted minimal-effort condition contains five.

Twisted-state formation. Both permuted medium-effort ring trials started with low local order ($\overline{Z}(0) = 0.141$ and 0.117) and ended with $\overline{Z}_{\text{steady}} \approx 0.859$. Both satisfied the twisted-state classifier. One trial ended with winding number -1, opposite to the ordered ramp, indicating that twisted spatial structure can form rather than merely persist from the original ordered initialization.

The collapse detector and regime classifier measure different properties. One trial reached the collapse criterion at turn 16; the other was right-censored despite satisfying the twisted-state classifier at the end of the experiment. The ordered-initialization collapse times therefore should not be interpreted as times required to form twisted states from disorder.

Topology and initialization. Under ordered initialization, all eight minimal-effort ring trials were classified as chimera-like, while all three regular-graph trials synchronized. Under permuted initialization, the five ring trials produced two chimera-like, two synchronized, and one incoherent outcome; all five regular-graph trials synchronized.

Thus, regular-graph synchronization persists across both initializations, whereas the frequency of chimera-like ring outcomes depends on the starting configuration. The small and unequal sample sizes limit inference about regime frequencies.

Fitted-map sensitivity. Under ordered initialization, the fitted map predicted twisted outcomes in 12 of 22 evaluated conditions, compared with 2 of 22 under permuted initialization. For the medium-effort ring condition, the permuted map predicted a mixture of twisted and synchronized outcomes, whereas both observed LLM trials ended twisted. This limited comparison illustrates the initialization sensitivity of the fitted approximation rather than establishing quantitative agreement with the LLM dynamics.

B Theoretical Derivations and Numerical Model

B.1 Local-Order Convention Details

We follow Abrams & Strogatz (2004) in computing Z_i over $\mathcal{N}_i \cup \{i\}$ (the "self-included" convention). The robustness summary in §4 compares this to an alternative convention that excludes self and normalises by $2R$; the regime classification and ΔZ ordering are preserved under the swap. Per-cell shifts in $\bar{Z}_i$ values under the swap are tabulated in the release notebook.

Spatial contiguity. A retrospective audit of the 55 chimera-like trials found contiguous low-order domains of at least two agents in 53 trials (median longest run: 10 agents). The two exceptions were isolated, threshold-marginal agents. Both chimera-like trials under permuted initialization contained extended low-order domains (9 and 11 agents). Thus, the operational classifier generally

identifies spatially extended heterogeneity, without requiring the stricter dynamical criteria of classical chimeras.

B.2 Derivations for the collective dynamics

Twisted fixed points. For $\theta_i = 2\pi ki/N + c$, the self-inclusive neighbourhood sum is

$$\sum_{j\in\mathcal{N}_i\cup\{i\}} e^{\mathrm{i}\theta_j} = e^{\mathrm{i}\theta_i}\left(1 + 2\sum_{m=1}^{R}\cos\frac{2\pi km}{N}\right) = S_k e^{\mathrm{i}\theta_i}.$$

When $S_k > 0$, its direction is θ_i, so the noiseless averaging increment is zero. Linearizing the circular mean gives the multipliers in Eq. (5). For C_{50}^R, the one-twist remains stable through $R = 16$ and loses stability at $R = 17$.

Spectral contraction. Let v_m be a Laplacian eigenvector with eigenvalue λ_m. From Eq. (3),

$$Wv_m = \left(1 - \frac{\beta\lambda_m}{d+1}\right)v_m.$$

The consensus mode has multiplier one and every other mode contracts. The exact off-consensus factor is

$$\rho_\perp = \max\left\{1 - \frac{\beta\lambda_2}{d+1}, \left|1 - \frac{\beta\lambda_N}{d+1}\right|\right\}.$$

If $\beta(\lambda_2+\lambda_N) \le 2(d+1)$, as in all regular-graph conditions studied here, the Fiedler mode is slowest. Approximating $(1-\gamma)^t$ by $e^{-\gamma t}$ then gives Eqs. (6)–(7).

Noise-supported disagreement. For independent additive noise of variance σ^2, the stationary squared disagreement is approximately

$$F \approx \frac{\sigma^2(d+1)}{2\beta}\sum_{m=2}^{N}\lambda_m^{-1}.$$

Small nonzero eigenvalues therefore predict both slow contraction and greater noise-supported disagreement.

B.3 Measured update-rule parameters

Model	Effort	β	R^2	σ	q
qwen-2.5-72b	def.	0.118	0.042	0.278	15.5%
gpt-5-nano	min.	0.141	0.050	0.290	14.0%
llama-4-scout	def.	0.167	0.054	0.195	8.4%
gpt-5-mini	min.	0.221	0.089	0.287	6.6%
mistral-large-2512	def.	0.244	0.081	0.369	13.4%
mistral-large-2411	def.	0.336	0.093	0.422	9.5%
gpt-oss-120b	min.	0.340	0.197	0.234	1.9%
deepseek-v4-flash	def.	0.354	0.233	0.133	2.4%
gemini-3.1-flash-lite	low	0.411	0.241	0.114	1.7%
gpt-5-mini	low	0.468	0.771	0.043	1.4%
gpt-5-mini	med.	0.489	0.705	0.048	1.0%
gemini-3.1-flash-lite	med.	0.589	0.488	0.104	2.3%
gpt-oss-120b	def.	0.626	0.769	0.071	0.0%
gemini-3.1-flash-lite	def.	−0.043	0.002	0.202	1.7%

Table 4: Measured update-rule parameters across all 14 model/effort conditions, ordered by β. Origin-constrained OLS of each agent's move on its neighbour-mean offset, ring runs only, drive $|\cdot| \ge 0.05$ rad; σ is the transient noise scale in radians and q the verbatim-copy rate. Equal-weight reference ($\beta = 0.5$).

B.4 THE NUMERICAL PHASE DIAGRAM IN DETAIL

We simulate the full stochastic map on each graph type over a grid of (β, σ, q) at $T = 30$, classify each cell by the criteria of Section 4.3, and compare the predictions with 22 measured clock-task condition–radius cells (Figure 5). The failures on the chimera-like conditions are systematic: every one occurs at low β and high σ, where the linear approximation underlying the fitted map is weakest. These are the conditions whose real updates involve large circular jumps that Gaussian noise cannot produce.

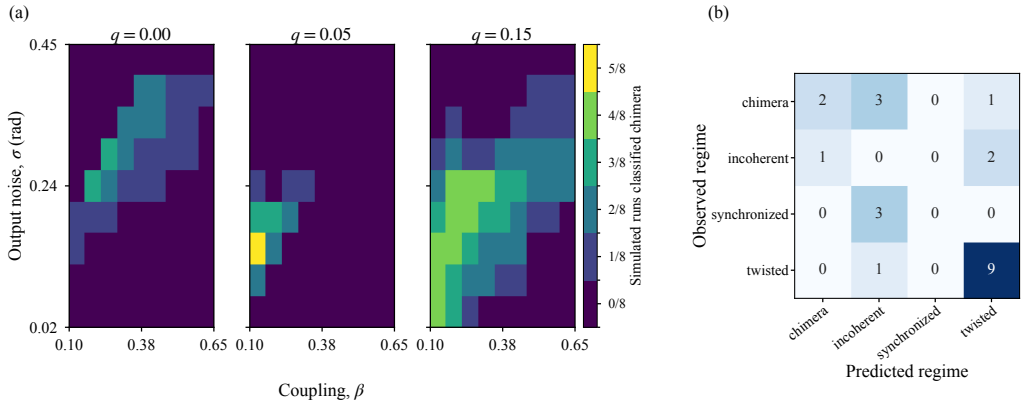

Figure 5: Fitted stochastic-map audit. (a) Fraction of simulated runs classified as chimera-like over the (β, σ) grid on a ring at $R = 7$ for three copying probabilities q (eight runs per grid cell). Chimera-like cells occur in 0/81 grid cells at $q = 0$, 1/81 at $q = 0.05$, and 9/81 at $q = 0.15$. (b) Observed versus predicted regimes for 22 evaluated condition–radius cells, representing 113 trials. With measured q, the map achieves 11/22 four-class accuracy, including 9/10 twisted and 2/6 chimera-like cells; it never predicts the synchronized class. These grid results do not establish that copying is generally necessary for chimera-like coexistence.

C SUPPLEMENTARY EXPERIMENTAL RESULTS

C.1 RADIUS-SWEEP RESULTS

	Qwen-2.5-72b			**gpt-5-nano**	
R	ΔZ	r	regime	ΔZ	regime
3	0.263	0.13	mixed	0.450	twisted (2/3)
4	0.628	0.37	chimera-like	0.273	twisted
5	0.630	0.45	chimera-like	0.483	chimera-like
6	**0.725**	0.36	chimera-like	**0.573**	chimera-like
7	0.543	0.26	chimera-like	0.555	chimera-like
9	0.562	0.52	chimera-like	0.398	mixed
10	0.484	0.62	chimera-like (2/3)	0.190	incoherent (2/3)

Table 5: Canonical steady-window ΔZ against coupling radius R, $n = 3$ per radius; bold marks each model's peak. Labels summarize trial-level regimes and report the majority where cells are mixed.

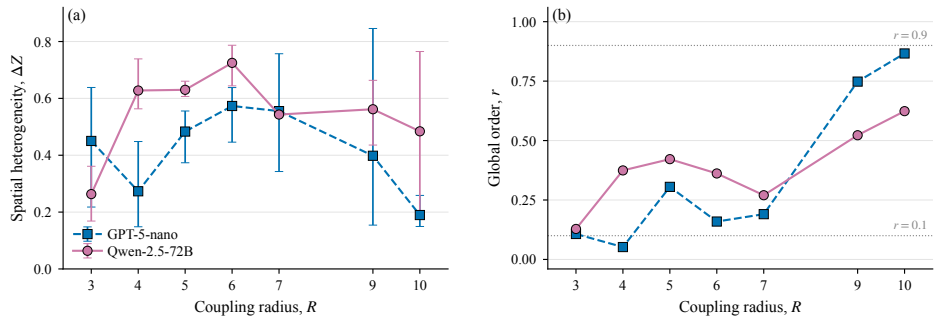

Figure 6: Spatial heterogeneity and global order across the $T = 30$ radius sweeps ($n = 3$ per radius). Both models attain highest observed mean heterogeneity at $R = 6$; the high-radius responses differ.

C.2 Model replication and cross-model analyses

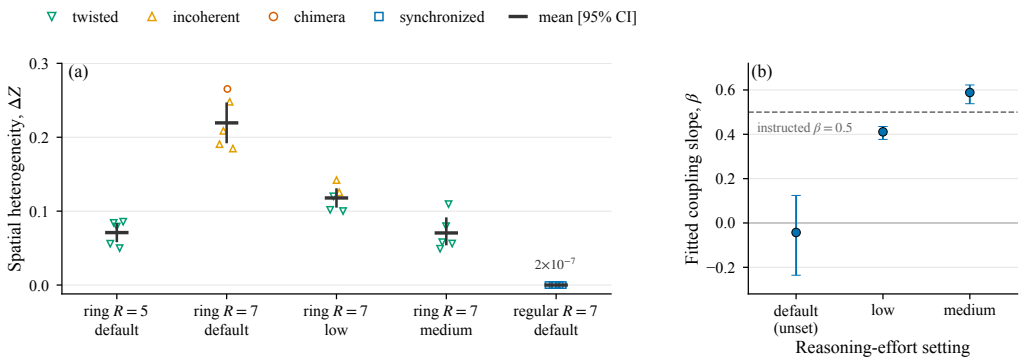

Figure 7: Gemini replication (five trials per condition; $N = 50$, $T = 30$, temperature 0). (a) Trial-level regimes and mean spatial heterogeneity with 95% confidence intervals. (b) Ring-only fitted coupling slopes for default (unset), low, and medium reasoning-effort settings; the dashed line denotes the equal-weight reference ($\beta = 0.5$). Default pools ring radii $R = 5$ and $R = 7$, whereas low and medium use only $R = 7$. Provider routing was not pinned and differed across conditions, so differences cannot be attributed exclusively to reasoning effort.

C.3 Network topology and task transfer

The judges-task primary analysis retains a pooled minimal-versus-medium reasoning-effort difference ($p = 0.0006$); the other two pooled effort contrasts are not significant. Passage-specific contrasts are exploratory. The topology results and full protocol-validity analysis are reported in the main text and Appendix A.1.

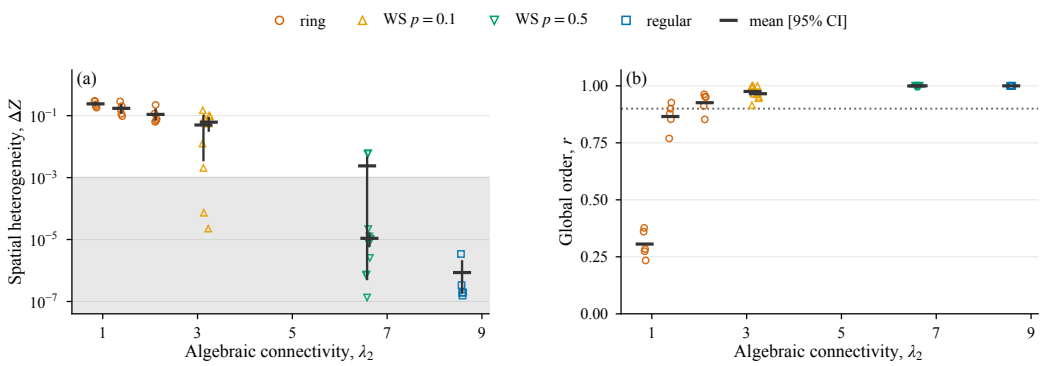

Figure 8: Watts–Strogatz spectral ladder for minimal-effort `gpt-5-mini` ($n = 5$ per graph). As measured algebraic connectivity λ_2 rises, canonical heterogeneity declines and global order approaches consensus. Distinct graph draws are retained rather than averaged.

D SPECTRAL AND PERSISTENCE ANALYSES

D.1 SPECTRAL APPROXIMATION AUDITS

Watts–Strogatz degree variation and contraction rates. The realized Watts–Strogatz graphs preserve mean degree exactly but not individual degrees. At $N = 50$ and $\langle k \rangle = 14$, the degree standard deviation is 1.0–1.1 at $p = 0.1$ and 2.5–2.6 at $p = 0.5$. For $P = (D + I)^{-1}(A + I)$ with eigenvalues μ_m, the update operator $W_\beta = (1 - \beta)I + \beta P$ has contraction factor off consensus $\max_{m \geq 2} |1 - \beta + \beta\mu_m|$. In the regime where the second eigenvalue controls contraction, the corresponding rate is $\gamma_{\text{true}} = \beta(1 - \mu_2)$, compared with the regular-graph approximation $\gamma = \beta\lambda_2/(\bar{d} + 1)$. The common factor β cancels in their relative discrepancy. On the realized instances the discrepancy is approximately 1% at $p = 0.1$ and 10% at $p = 0.5$, and zero on the ring and degree-preserving control.

Twisted-sector rate comparison. The slowest non-consensus multiplier of Eq. (5) gives a rate distinct from $\beta\lambda_2/(d+1)$. The two rates agree to within 6%, 9%, and 13% at $R = 5, 6, 7$, respectively, the radii used in the lifetime experiments. The discrepancy reaches 27% at $R = 10$. These comparisons motivate treating $1/\lambda_2$ as an approximate scaling prediction for the twisted sector rather than an exact contraction law.

Finite-size spectral comparison. The asymptotic Friedman bound has an ε allowance and is not a guaranteed spectral floor at finite N (Friedman, 2008). The realized random-regular controls have $\lambda_2 = 4.96$ for $d = 10$ and 8.58 for $d = 14$, exceeding the asymptotic reference values 4.00 and 6.79. Table 6 uses measured gaps at $R = 5, 7$ and the asymptotic reference elsewhere. At $R = 7$, the ring modes $m = \pm 1, \pm 2$ lie below the measured gap; only $m = \pm 1$ lie below the asymptotic reference. The full ring spectrum is checked at each swept radius: no additional interior mode falls below the corresponding comparison gap.

D.2 GRAPH SPECTRAL QUANTITIES

Spatial structure of surviving fragments. We compare the ring spectrum with the spectral gap of the random-regular control. For the $N = 50$ graphs used here, the ring modes below this gap are the lowest spatial harmonics, $m = \pm 1$ and, at smaller coupling radii, $m = \pm 2$ (Table 6). Since the ring spectrum is not monotonic in $|m|$, this comparison is made against the full spectrum rather than the lowest modes alone.

Graph (d)	λ_2	$(d+1)/\lambda_2$	$n_<$
Ring R=3 (6)	0.22	32.0	4†
Ring R=4 (8)	0.47	19.3	4†
Ring R=5 (10)	0.85	13.0	4
Ring R=6 (12)	1.39	9.3	4†
Ring R=7 (14)	2.12	7.1	4
Ring R=9 (18)	4.19	4.5	2†
Ring R=10 (20)	5.57	3.8	2†
Random reg. (10)	4.96	2.2	0
Random reg. (14)	8.58	1.7	0

Table 6: Spectral quantities at $N = 50$. Dividing $(d+1)/\lambda_2$ by the measured β gives the characteristic collapse time τ. Random-regular rows report measured gaps. $n_<$ counts ring modes below the degree-matched control gap, using measured gaps at $R = 5, 7$ and the asymptotic reference elsewhere (†).

D.3 LIFETIME CALIBRATION

The closed-form approximation overpredicts collapse time at all seven observed cell means (by 1.13–3.18 times), but not at every individual trial: 33 of 35 observed collapses lie below the identity line and two lie above. Five additional trials are right-censored. The main-text spectral-lifetime figure presents the trial-level comparison.

D.4 PERSISTENT-MINORITY AUDIT

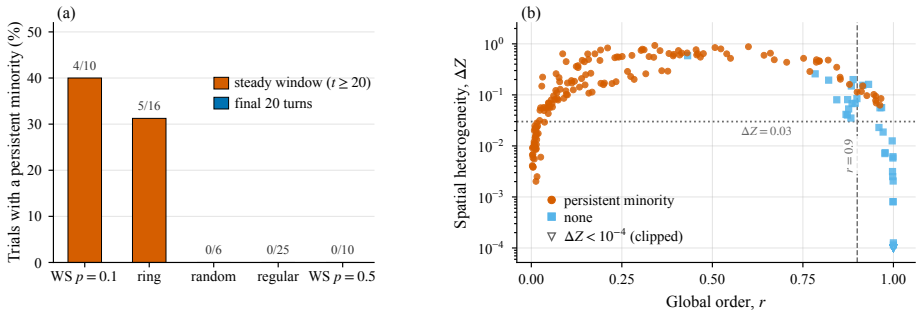

Figure 9: Persistent-minority audit. (a) Among the 67 LLM clock trials with steady-window global order $r \geq 0.9$, persistent minorities occur in 4/10 WS $p = 0.1$ and 5/16 ring trials during the steady window ($t \geq 20$), and in none of the trials during the final 20 turns. (b) All 228 LLM clock trials, with each point classified using the same steady-window persistent-minority detector as in (a). The vertical line marks $r = 0.9$ and the horizontal line marks $\Delta Z = 0.03$; values below 10^{-4} are clipped for display. Low spatial heterogeneity alone does not imply global synchronization: all 36 trials with $r < 0.9$ and $\Delta Z < 0.03$ are classified as twisted. Panel (a) aggregates a filtered subset by topology, whereas panel (b) shows individual trials across the full clock-task set. Both exclude non-LLM calibration baselines.