PNAS RESEARCH ARTICLE | APPLIED MATHEMATICS SOCIAL SCIENCES
OPEN ACCESS

Group size effects and collective misalignment in LLM multi-agent systems

Ariel Flint[a], Luca Maria Aiello[b,c], Romualdo Pastor-Satorras[d], and Andrea Baronchelli[a,1]

Edited by Giorgio Parisi, Universita degli Studi di Roma La Sapienza, Rome, Italy; received November 5, 2025; accepted July 14, 2026

Multi-agent systems of large language models (LLMs) are rapidly expanding across domains, introducing dynamics not captured by single-agent evaluations. Yet, existing work has mostly contrasted the behavior of a single agent with that of a collective of fixed size, leaving open a central question: How does group size shape dynamics? Here, we move beyond this dichotomy and systematically explore outcomes across the full range of group sizes. We focus on multi-agent misalignment, building on recent evidence that interacting LLMs playing a simple coordination game can generate collective biases absent in individual models. First, we show that collective bias is a deeper phenomenon than previously assessed: Interaction can amplify individual biases, introduce new ones, or override model-level preferences. Second, we demonstrate that group size affects the dynamics in a nonlinear way, revealing model-dependent dynamical regimes. Finally, we develop a mean-field analytical approach and show that, above a critical population size, simulations converge to deterministic predictions that expose the basins of attraction of competing equilibria. These findings establish group size as a key driver of multi-agent dynamics and highlight the need to consider population-level effects when deploying LLM-based systems at scale.

size effects | multi-agent systems | large language models | collective misalignment | social coordination

Significance

Large language models (LLMs) are increasingly deployed in large numbers, and their interactions make collective behavior harder to anticipate than that of a single model. While most studies compare one model with a collective of fixed size, we ask a key yet overlooked question: What is the role of group size? We show that interaction among LLMs can magnify individual biases, generate new ones, or even overturn individual preferences, and that, crucially, these effects scale in unexpected, nonlinear ways with group size. Our results demonstrate that more is different for LLM populations: The number of interacting agents is a key driver of the dynamics, with implications for the design and governance of multi-agent AI systems.

Multi-agent large language model (LLM) systems have recently started to be deployed and are already responsible for tasks in domains such as finance (1, 2), defense (3), energy (4), social media (5), and personal assistance (6). These applications are expected to expand rapidly, as reflected in growing market interest, with valuations projected to increase from USD 5.4 billion in 2024 to USD 236 billion by 2034, and a surge in startups and enterprise features focused on interacting LLM agents (7, 8). Scalability, autonomy, and interaction, which are at the root of the success of LLM-based multi-agent systems, also give rise to new risks. Building on evidence that multi-agent safety is not guaranteed by single-agent safety (9), the International AI Safety Report 2025, authored by 100 experts including representatives from 33 countries and intergovernmental organizations, has warned of the unprecedented complexity of multi-agent LLM systems (10), highlighting risks such as systemic failures, misaligned collective behaviors, and uncontrollable strategic dynamics (11). As multi-agent LLM systems become a dominant paradigm in AI infrastructure, understanding their dynamics, including both capabilities and failure modes, is widely recognized as a central scientific and governance priority (12, 13).

Recent research has shown that local interactions among agents can yield collective behaviors and performance gains that are not observed when models operate in isolation (14, 15). For example, at the smallest scale, dyads (pairs of agents) engaging in debate lead to substantial improvements in factual accuracy and truthfulness (16), and may also enhance coherence and depth of LLM answers (17). Triads of LLMs have been shown to produce more utilitarian collective moral judgments than individual models (18), and may optimize collaboration in complex tasks (19). Small groups of agents can accomplish goals more effectively through self-organized division of labor (20). Groups of tens of agents are able to develop social conventions (21), and to autonomously coordinate social behaviors that mirror aspects of human interaction in immersive environments (22). Populations on the order of hundreds or thousands have been used to investigate mechanisms of cultural evolution (23–25), while even larger populations, scaling from tens to hundreds of thousands, have been studied as in-vitro societies (26). Frameworks have also been proposed to model the dynamics of millions (27), or even billions (28), of interacting LLMs.

While convincingly supporting the idea that "more is different" for populations of LLMs (29), the vast majority of studies in this rapidly growing body of work are

Author affiliations: [a]Department of Mathematics, City St George's, University of London, London EC1V 0HB, United Kingdom; [b]Data Science Section, IT University of Copenhagen, Copenhagen 2300, Denmark; [c]Networks and Graphs Collaboratory, Pioneer Centre for AI, Copenhagen 1350, Denmark; and [d]Departament de Física, Universitat Politècnica de Catalunya, Barcelona 08034, Spain

Author contributions: A.F., L.M.A., R.P.-S., and A.B. designed research; performed research; contributed new reagents/analytic tools; analyzed data; and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: a.baronchelli.work@gmail.com.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2531697123/-/DCSupplemental.

Published August 18, 2026.

characterized by a recurring methodological structure: The behavior of a population is compared with that of a single agent, revealing differences between the two conditions. However, this approach leaves a fundamental question unanswered: What happens across varying population sizes? In other words, what does "more" quantitatively mean? This question is crucial for applications involving scalable deployments of LLM agents, where it is critical to understand whether there exists a threshold beyond which increasing the population size no longer affects system behavior, whether in terms of performance, coordination, or emergent properties.

Here, we investigate the role of group size on LLM collective dynamics by focusing on bias, a critical risk associated with language models that has been extensively studied in isolated settings and carries far-reaching implications for safety and alignment (30–33). We build on recent empirical observations showing that groups of LLMs coordinating on shared linguistic conventions can exhibit collective misalignment (21), introducing bias even when individual agents are unbiased. This setting deliberately abstracts away from contextual grounding, in contrast to context-rich benchmarks (34), to isolate the role of interaction dynamics in shaping collective bias. In this context, individual bias refers to an initial statistical preference for one option over an equivalent alternative during coordination (for example, agents systematically preferring one name over another in a naming process). Collective bias, by contrast, denotes an additional imbalance that emerges from interactions among multiple LLMs, increasing the likelihood that certain coordination outcomes are adopted over others that are, in principle, equally viable. Thus, we use bias as a tractable operationalization of collective misalignment, focusing on cases where collective outcomes diverge from individual-level preferences. This notion is internal to the coordination process and does not directly correspond to misalignment with external human values as typically considered in the AI safety literature (9, 35).

The contribution of this paper is threefold. We clarify and extend the phenomenon of collective bias as a form of collective misalignment, showing how coordination can amplify, create, or overturn individual biases. We map how the strength and form of collective bias depend on group size, identifying nonlinear effects and qualitative shifts in the dynamics. We provide an analytical framework that clarifies why simulations converge beyond a given population size and how basins of attraction structure the competition between equilibria. The findings are robust across different LLMs. These results are obtained within a simplified and controlled framework, which distinguishes our approach from goal-oriented multi-agent LLM studies focused on task completion or human behavior simulation (14, 22, 36). Adopting a complex systems perspective (13), we prioritize mechanistic transparency and use minimal models to isolate the fundamental dynamics driving collective outcomes (21, 37).

Modeling Framework and Parameterization

Modeling Framework. We consider a population of N LLM agents engaged in a naming game (38, 39), a standard framework for studying the emergence of conventions that has been extensively investigated through theoretical modeling (37, 40) and laboratory experiments with human participants (41, 42). In particular, we follow the implementation used in the latter, where pairs of agents drawn from a larger population attempt to coordinate through simultaneous naming choices, receive rewards for successful coordination, and, after each interaction, observe both their own and their partner's choice.

At each time step, two agents are randomly selected to interact, producing and exchanging a convention, or word, chosen from a finite pool of size W, with the goal of maximizing their respective game scores (see *Materials and Methods* and *SI Appendix* for prompting and meta-prompting details). The outcome of an interaction yields identical payoffs to both agents. If both agents select the same convention, their scores increase by the same amount; otherwise, they decrease by the same amount. The reward for successful coordination exceeds the penalty for failure. This scoring rule incentivizes local coordination between pairs of agents, without any explicit drive toward global consensus. Each agent maintains a finite memory state M of capacity H, which stores information about its most recent interactions: Its own and its partner's convention choices, whether coordination was achieved, and the accumulated score over the last H encounters. In this work we fix $H = 5$. At the beginning, memories are empty, and the first choice is drawn from the available pool according to the agent's individual bias. Recent analyses of this protocol in LLM populations reveal that such local interactions can spontaneously lead to global consensus on a specific word and, crucially, generate collective bias even when individual agents exhibit none (21).

This minimal signaling framework can be extended to interaction settings in which agents communicate short natural-language responses during each encounter, for example by providing brief justifications for their convention choices that are incorporated into interaction memory. Such extensions increase the expressive dimensionality of the signaling channel while preserving the underlying coordination incentives of the naming-game protocol. We examine this conversational variant in *SI Appendix* and find that it preserves the key size-dependent dynamics observed in the minimal framework. While we consider this an important robustness check, we retain the minimal framework as the primary setting since the richer interaction protocol entails substantially higher computational cost and limits both the range of population sizes that can be explored and the statistical resolution achievable. In addition, it increases the dimensionality of agent policies and interaction histories, making it more difficult to isolate coordination mechanisms and to connect the observed dynamics to the analytical mean-field framework developed here.

Implementation. The modeling framework described above can be directly implemented by running LLMs as agents, where the memory state of each agent is translated into a text prompt that the LLM uses to generate a new convention (21). To study the coordination dynamics of large populations of LLM agents, however, we introduce a complementary stochastic model that enables simulations with population sizes far beyond what is computationally feasible in direct LLM experiments, while drastically reducing the required computational resources.

In this stochastic model, an agent's behavior is represented by a probabilistic policy that specifies the likelihood of producing each of the W possible words based on the agent's current memory state. We derive this policy empirically by probing the LLM: Given a textual prompt encoding a particular memory state, the LLM outputs a vector of logits indicating the likelihood of all possible next tokens. We extract and normalize the logits corresponding to our restricted set of W words to form a probability distribution, which defines the agent's policy for that memory state (see *Materials and Methods* for full details on model definition and policy extraction procedure).

To enable efficient large-scale simulations, we precompute and store these probabilistic policies for all possible memory states. During simulation runs, agents sample conventions according to their stored policies, thereby capturing the stochastic nature of LLM-generated interactions without invoking the underlying model at every step. This approach is validated by extensive experiments showing close agreement between direct LLM-based generations and policy-driven simulations (see *SI Appendix* for quantitative comparisons, particularly *SI Appendix*, Figs. S1 and S2).

Parameterization. In this paper, we restrict our analysis to the case of $W = 2$ competing conventions. Since we use multi-agent bias as the main measurement to explore size effects, we select pairs of words previously identified as particularly sensitive to bias (e.g., {*man, woman*} or {*straight, gay*}) (34). To test robustness across architectures, we feed the model with probability policies extracted from different instruction-tuned LLMs: Qwen QwQ-32B (denoted as Qwen), Microsoft Phi-4 (Phi), OpenAI GPT-4o (GPT), and Meta Llama 3.1 Instruct (Llama).

We consider only homogeneous populations, meaning that the probabilistic policies of all agents are extracted from the same underlying LLM. Finally, throughout the paper, time is measured in population rounds (Monte-Carlo steps), each consisting of N pairwise interactions. Simulations run for up to 1,000 population rounds or until all agents converge on the same convention. Theoretical predictions for simulations are based on the minimal naming game model (39), in which agents may additionally be individually biased toward one convention (*SI Appendix*).

Results

Mapping Collective Misalignment. We begin by examining how individual bias shapes collective outcomes by performing simulations of our model for small system sizes. Fig. 1 shows that individual biases are not necessarily reflected by collective outcomes. Instead, interaction gives rise to three distinct forms of misalignment between individual and collective bias. First, individual preferences may be amplified (Fig. 1 *A* and *D*): Collective dynamics increase the likelihood of consensus on the word agents initially favor. Second, neutral individuals can give rise to collective bias, consistent with previous findings (21) (Fig. 1 *B* and *E*). Finally, collective dynamics can override individual preferences, producing consensus on the word agents initially disfavor (Fig. 1 *C* and *F*).

To systematically assess the robustness of these misalignment patterns, we extend the analysis across eleven word pairs and four homogeneous LLM populations. Fig. 2 contrasts individual and collective biases for four LLMs in populations of size $N = 24$, revealing strong variations across LLMs and word pairs. The dashed black line indicates theoretical predictions from the minimal naming game model, which assumes agents with fixed, memory-invariant bias. LLM agents, by comparison, adjust according to the memory context. It is worth noting that the minimal naming game captures only the amplification of individual bias, while induction and reversal represent qualitatively distinct outcomes that, to the best of our knowledge, have not been observed in standard naming-game-like models and are genuinely LLM-dependent in this context.

At the individual level, where single agents have empty memory states prior to interaction, Phi and Llama display considerable variability in individual bias, whereas GPT and Qwen are closer to neutrality. At the collective level, however, coordination outcomes differ markedly across models. Qwen populations exhibit a wide spectrum of collective bias strengths, GPT populations show only mild collective effects, while Llama populations are highly polarized and converge almost deterministically on a single word for each pair. Phi populations most closely align with the minimal naming game model predictions, though they still diverge strongly in five of the eleven word pairs.

Importantly, both the magnitude and the direction of collective bias can differ across models. For example, for the pair {*her, his*}, populations of Qwen and Phi converge on *her*, while GPT and Llama converge on *his*, despite nearly identical individual-level tendencies. Moreover, bias induction and its strengthening with population size are also observed under the more complex signaling mechanism, while further testing is needed to map the full spectrum of outcomes, including reversal (*SI Appendix*, Fig. S3).

The Role of Population Size. Having established that LLM populations can exhibit diverse collective behavior, we next examine the role of population size in shaping these outcomes. Fig. 3 shows that the probability of converging on a specific word grows systematically with N across all models and word pairs. In other words, larger populations reduce outcome uncertainty. We label the preferred word as the "strong" word, as opposed to its "weak" alternative. In the minimal naming game with individually biased agents, such a transition to determinism is indeed expected (43): Bias amplification causes the favored convention to fully dominate collective outcomes beyond a critical population size (*SI Appendix*, Fig. S4). In LLM populations, collective outcomes also become entirely deterministic once N exceeds a threshold: If the population can reach consensus, then it will always converge on the strong word. This pattern of increasing collective bias and eventual determinism in consensus outcomes is also observed in the richer signaling setting (*SI Appendix*, Fig. S3).

The threshold varies considerably across model and word pair combinations. In some cases, determinism arises for populations as small as $N = 2$ (Llama, {*short*, *tall*}), while in others, finite-size effects persist up to $N \sim 10^4$ (Qwen, {*Black*, *White*}). The rate at which collective bias grows with N likewise depends on both the LLM and the word pair. Fig. 3*C* shows that although Qwen exhibits uniform strategies at the individual level ($N = 1$) for all word pairs, the magnitude of collective bias at intermediate N values varies. Population size can also modulate the form of multi-agent misalignment: For the word pair {*straight*, *gay*}, Llama shows an individual preference for *straight*, but reversal toward *gay* occurs only for $N \geq 6$. Finally, we note that populations always converge, except large Llama populations for the word pairs {*old*, *young*} and {*less*, *more*}, and a minor fraction of simulation runs for small Qwen populations coordinating on the word pair {*husband*, *wife*}. We document these in *SI Appendix* and defer systematic analysis to future work (*SI Appendix*, Figs. S5–S9).

Convergence Time. The dependence of collective bias on population size can be explained by qualitative shifts in the underlying consensus dynamics. In small populations, early random fluctuations dominate the dynamics. A word that gains even a small initial advantage is quickly incorporated into the memory inventories of most agents, driving rapid alignment across the system. As shown in Fig. 4, trajectories collapse to consensus (on either word) within only a few population rounds. Because the

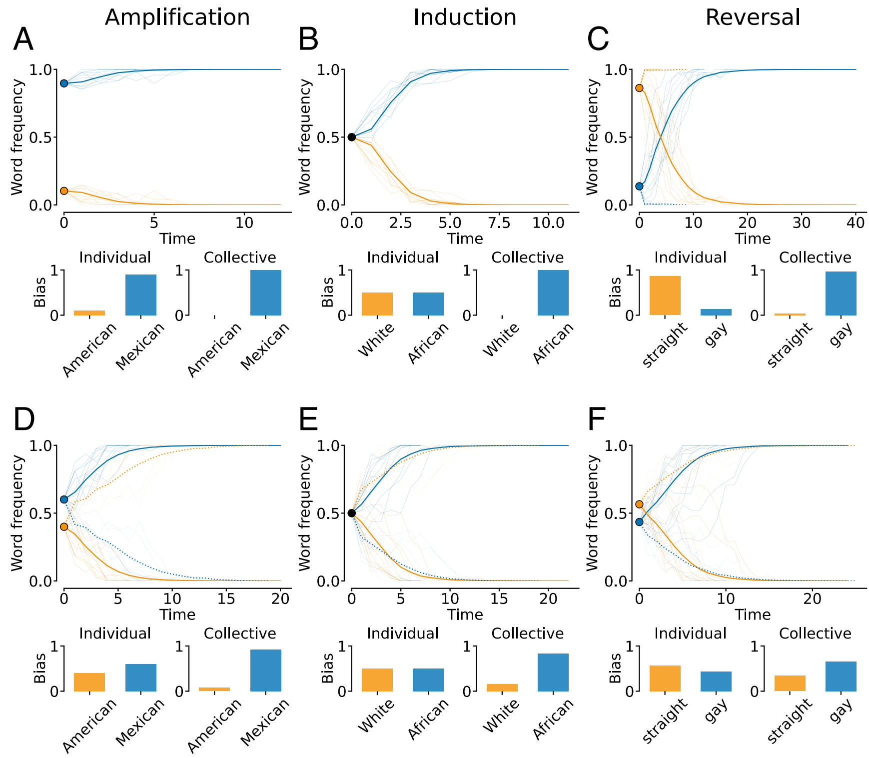

Fig. 1. Interaction can amplify, induce, or override individual bias. From *Left* to *Right* columns, the word pairs that populations coordinate on are {*American, Mexican*}, {*White, African*}, and {*straight, gay*}. These cases illustrate bias amplification, induction from neutrality, and bias reversal, respectively. *Top* row (*A–C*): Llama populations; *Bottom* row (*D–F*): GPT populations. All simulations use a population size of $N = 24$. The *Upper* plots in each panel display representative trajectories of word competition in 1,000 simulation runs, where blue and orange lines indicate the frequency of unique convention choices over time (based on the previous N interactions). Colored circles at $t = 0$ denote the initial individual bias, with black indicating neutrality. Where possible, at least ten trajectories are shown for each consensus outcome (strong or weak convention); when fewer runs converged, all available trajectories are displayed. Solid and dotted lines show the mean dynamics of runs that converged on the strong and weak conventions, respectively. In all cases, the strong word is represented by blue trajectories (from *Left* to *Right*: *Mexican*, *African*, *gay*). The *Lower* bar plots summarize individual and collective bias: Individual bias reflects agents' preinteraction preferences, and collective bias shows the fraction of runs that reached consensus on each convention.

likelihood of a word becoming an early favorite depends on the agents' initial individual biases, the collective outcome ultimately reflects these underlying preferences.

In intermediate size populations, initial fluctuations are rarely sufficiently strong to produce the rapid alignment observed at small N. Consensus arises from the interplay of stochastic fluctuations, which may affect the prevalence of one word even beyond the initial phase, and coordination dynamics, which tends to favor the strong word, steering trajectories toward its absorbing state. This coexistence of rapid collapses and prolonged mixed states produces a unimodal distribution of consensus times with a heavy right tail (Fig. 4). The crossover to coordination-dominant dynamics is also reflected in the scaling of the characteristic (i.e., modal) consensus time, which grows logarithmically with population size (*SI Appendix*, Figs. S10–S13 and S14), consistent with the coordination-driven scaling of the minimal naming game (43) (*SI Appendix*, Fig. S15).

As populations grow even larger, the consensus time distributions separate: The characteristic consensus time for trajectories converging on the weak word becomes larger than the consensus time to the strong word (*SI Appendix*, Figs. S10–S13). This widening gap suggests that it becomes increasingly difficult for the population to converge on the weak word. Beyond the threshold size N_c, the consensus state becomes deterministic: For a given word pair and model, all simulations converge on the strong word and the characteristic consensus time stabilizes (*SI Appendix*, Figs. S16–S19). The consensus time distribution for trajectories that converge on the strong word (the only viable consensus state) narrows, until N is so large that nearly all trajectories converge at this characteristic convergence time.

The coordination dynamics depend on the model and the word pair. In practice, this means that while the qualitative shift from fluctuation-driven to coordination-driven consensus is robust, the transformation stages of the consensus time PDF can differ in duration and form across experimental conditions (*SI Appendix*, Figs. S20–S30).

Mean-Field Theory. To understand the deterministic outcomes observed in large populations, we develop a mean-field theory that captures the system dynamics in the limit $N \to \infty$. Mapping the dynamics of the experimental framework to a reaction–diffusion process (*SI Appendix*), we define the system state as $\mathbf{x} := \{x_i\}$, where $x_i = N_i/N$ represents the fraction of agents in state i at a given timestep. The evolution of the system is characterized by the following mean-field rate equation:

$$\frac{dx_k}{dt} = -x_k + \sum_i \sum_j x_i x_j P_k(i,j), \quad [1]$$

where $P_k(i,j)$ is the probability that an agent in memory state i will transition to a new memory configuration k after interacting with an agent in memory state j. These dynamics are characterized by fixed points given by the algebraic equation

$$x_k = \sum_i \sum_j x_i x_j P_k(i,j). \quad [2]$$

case, one fixed point is unstable while the other is marginal, with an almost vanishing largest eigenvalue. A different situation arises for Qwen populations coordinating on the pair {*husband*, *wife*}. Here, the weak fixed point is marginal but the strong one is stable. Consequently, large populations converge reliably to the strong consensus state, while in smaller populations convergence toward the weak word remains susceptible to random fluctuations, which can prevent the system from settling into a stable consensus on that word.

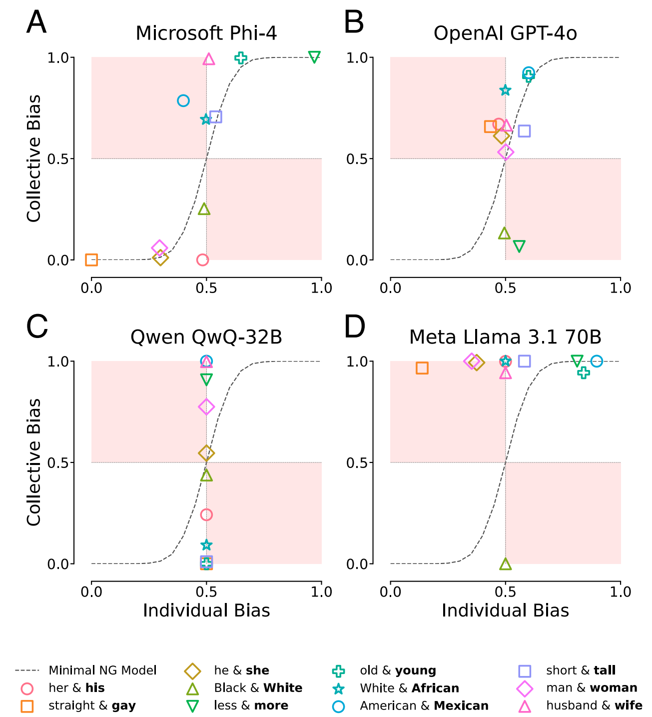

Fig. 2. Collective bias depends on both model and convention. Panels show individual and collective bias for four LLMs [Panels (*A*–*D*) show Phi, GPT, Qwen, and Llama populations, respectively] in populations of size $N = 24$, for the word pairs indicated in the legend. Individual bias corresponds to an agent's selection probability at the start of the game with empty memory, while collective bias denotes the proportion of 1,000 runs that reached consensus on the option in bold. Error bars (SEM) are smaller than the marker size. The dashed line shows the theoretical prediction from the minimal naming game with binary options and individual bias. Points along the gray dotted line at $x = 0.5$ indicate symmetry breaking, where neutral individual preferences produce biased collective outcomes. Points within the shaded pink regions correspond to bias reversal, where collective interactions overturn the individual preference.

These fixed points take the form of homogeneous absorbing states, corresponding to full consensus on one of the two available options. That is, the memory of all agents is equal to $2H$ repetitions of the first word, or $2H$ repetitions of the second word.

We therefore denote them as the strong and weak fixed points. Although, in principle, other, more complex mixed fixed points could exist, they are not generally observed in numerical solutions of the mean-field equations. Only for a few specific combinations of LLMs and word pairs (i.e. {*old*, *young*} and {*less*, *more*} in large Llama populations) do mixed configurations appear, corresponding to time-dependent steady states also observed in direct simulations of the agent-based model at finite N. The mean-field dynamics for all models are shown in *SI Appendix*, Figs. S31–S34.

A linear stability analysis allows us to determine if the homogeneous fixed points are stable or not by looking at the sign of the largest eigenvalue of the corresponding Jacobian matrix. In *SI Appendix*, Table S2, we report the largest eigenvalues computed for the weak and strong fixed points. As we can see, for most combinations of LLM and word pair either the strong fixed point is stable and the weak one unstable, or both are stable. In this last case, the actual preference for the strong word is due to the structure of the basins of attraction of the two fixed points.

Finally, this analysis explains the observed cases of nonconsensus for the word pairs {*less*, *more*} and {*old*, *young*} in Llama populations. In the first case, both homogeneous fixed points are unstable, so trajectories are repelled from consensus states and the population remains in a mixed configuration. In the second

Discussion

In this work, we have shown that interactions among LLM agents can lead to collective misalignment, with emergent biases whose strength and form vary systematically with population size. Crucially, each agent conditions its behavior on a finite memory of past interactions, so that collective outcomes emerge from the interplay of individual history-dependent policies and population-level dynamics. Our mean-field analytical framework captures these patterns, linking stochastic simulations to deterministic predictions and clarifying the structure of the equilibrium landscape.

The origin of bias in our setting deserves comment. Unlike context-rich frameworks such as CrowS-Pairs (34), where bias is triggered by sentence-level context, our minimal coordination task does not include external referential grounding. As a result, individual bias may reflect statistical regularities in pretraining

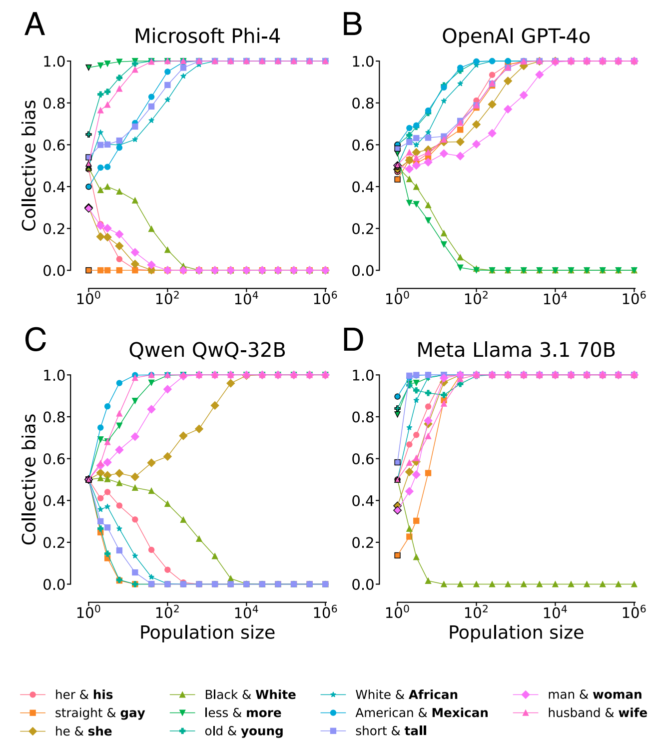

Fig. 3. Population size effects on collective bias. For all convention pairs tested, collective preference for a given convention increases with population size until convergence becomes deterministic. Panels (*A*–*D*) show Phi, GPT, Qwen, and Llama populations, respectively. The individual bias is reported at $N = 1$. Error bars (SEM) are smaller than the marker size. Across population sizes, all runs reached consensus except in three cases: In Llama populations, the convergence rate decreases beyond a certain N for {*old*, *young*} and {*less*, *more*}, and in small Qwen populations, a few runs failed to converge for the word pair {*husband*, *wife*}. See *SI Appendix*, Figs. S5–S9 for word competition dynamics and precise convergence fractions for these cases. *SI Appendix*, Table S1 shows the critical thresholds at which simulations reach deterministic outcomes.

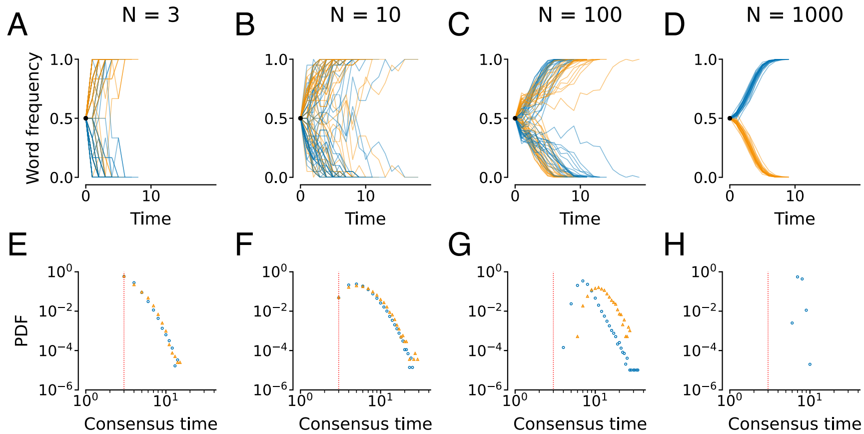

Fig. 4. Coordination dynamics become increasingly deterministic for larger populations. Each column corresponds to a different population size ($N =$ 3, 10, 100, 1,000) for GPT populations coordinating on the word pair {*White*, *African*}. Panels (*A*–*D*) show temporal trajectories from individual simulation runs (up to 25 per consensus outcome) of the usage fraction of the strong word (*African*, blue) and the weak word (*White*, orange). Panels (*E*–*H*) show the corresponding probability density functions (PDFs) of consensus times, that is, the time taken for a trajectory to converge, computed from 100,000 trajectories that reached consensus. Blue circles and orange triangles indicate convergence to the strong and weak words, respectively. The dotted red line marks the minimum possible consensus time under the convergence criterion ($t = 3$). Across columns, as N increases from *Left* to *Right*, trajectories increasingly favor the strong word and reach consensus on it more reliably, as reflected in both the clustering of trajectories in panels (*A*–*D*) and the sharpening of the consensus time PDFs toward convergence on *African*. The probability of coordinating on the strong word increases with population size (to three significant figures): 0.599, 0.720, 0.981, and 1.00.

data, although posttraining alignment procedures and prompt-level framing likely also play a role, with contributions that are difficult to disentangle in the present setting. Collective bias, however, cannot be reduced to a simple reflection of individual-level tendencies: Its pronounced model dependence, together with the fact that collective outcomes can depart from, and even invert, individual preferences, suggests that it emerges from the interplay of model-specific policies, training data statistics, and, crucially, the accumulation of interaction memory by agents.

The strong dependence of outcomes on population size establishes it as a critical determinant of behavior in multi-agent LLM systems rather than a neutral parameter. In the locally interacting systems examined here, this dependence plays a particularly prominent role, producing behavioral transitions—including induction and reversal regimes—that cannot be explained by simple one-versus-N comparisons and are not captured by the minimal naming game (39), where deterministic outcomes arise only through the amplification of preexisting individual bias. Related work has shown that populations of globally informed LLM agents fail to reach an ordered state beyond a certain size threshold (24). However, since these agents have access to the full population state, that setting is intrinsically distinct from ours, which focuses on how local interactions translate into collective, macrolevel outcomes. Expanding investigations to more local-interaction frameworks, for example to classic opinion dynamics models that account for minority opinion spreading (44), represents a promising direction for future research.

Further limitations of the present study, arising from our focus on isolating group size effects while limiting confounding factors, point to several directions for future work. First, we assume homogeneous populations: All agents are instantiated from the same LLM, and interactions occur under homogeneous mixing, where any agent can in principle interact with any other (i.e., a fully connected network). Natural extensions include embedding agents in more realistic interaction topologies, such as scale-free or structured networks, and considering heterogeneous populations composed of different LLMs, with the pronounced model dependence already observed here suggesting that the latter may exhibit particularly rich dynamics. Second, the word pairs examined here were selected because they carry known social meaning in LLM outputs, allowing us to test whether collective dynamics can reshape preferences along socially salient dimensions. While we make no normative claims about which collective outcomes are desirable, an open question is whether collective bias dynamics differ systematically across domains, for example whether convergence is faster or more deterministic along some axes than others, with implications for how multi-agent LLM systems are evaluated and governed. Finally, the individual-level policies underpinning our results are necessarily task-specific: Biases extracted in this minimal coordination setting may differ from those that would emerge under richer conversational framing or more naturalistic task contexts. Our framework deliberately holds the interaction structure fixed to isolate population size effects from such confounds, following the logic of minimal models widely used in the social and physical sciences. How context-dependent shifts in individual policies propagate to collective outcomes remains an open and important question.

Overall, our focus here has been on collective bias, a specific and tractable form of misalignment between individual-level preferences and collective decision outcomes, distinct from misalignment with external human values or intentions. Yet the implications of our findings extend beyond this setting: Given the results presented here, group size effects across interaction-driven phenomena such as collusion, deception, and cooperation can no longer be assumed negligible without empirical investigation. Current testing practices may overlook scale-specific risks that arise only at particular population sizes, whether intermediate or large, and this calls for a broader research agenda to understand size effects across tasks and domains. Advancing this agenda is crucial for developing reliable frameworks and tools to predict, control, and safely deploy complex collective behaviors in LLM societies, and for establishing multi-agent AI as a systematic scientific discipline.

Materials and Methods

Model Definition. The model is defined in terms of a population of N agents, playing the naming game with two word options, that we label as π_1 and π_2. At each time step, two agents are randomly selected for interaction. Each agent will then output one of the two available words, and exchange it with the other agent. Each agent has a finite memory of maximum length H, storing the words proposed and received in the last interactions. Thus, the memory state of an agent at any time can be written as

$$M_h(t) := \{a_1, b_1, a_2, b_2, \dots, a_h, b_h\}, \qquad [3]$$

with $h \leq H$, where a_i are the most recent words proposed by the agent and b_i are the most recent words received from interaction partners ($i = 1, \dots, h$). All agents start from the empty state $M_0 = \emptyset$, which corresponds to the memory before any interaction with peers. There are a total of $\mathcal{N}_H = \frac{1}{3}(4^{H+1} - 1)$ possible different states. An agent with a state M_h will propose an option π given by

$$\pi = \begin{cases} \pi_1 & \text{with probability } q(M_h), \\ \pi_2 & \text{with probability } 1 - q(M_h). \end{cases} \qquad [4]$$

The probability $q(M_h)$ that an agent in state M_h produces the word π_1 encodes the probability policy characteristic of the LLM it represents. See below for a description of how the probability $q(M_h)$ is extracted from the corresponding LLM.

In the coordination dynamics of the experimental setting, two agents are chosen at random, with states M_h and $M_{h'}$, respectively. During interaction, these agents propose options π and π' according to Eq. **4**. The states of the agents then change to

$$M_h \rightarrow \Gamma(M_h, \pi, \pi'), \qquad [5]$$

$$M_{h'} \rightarrow \Gamma(M_{h'}, \pi', \pi). \qquad [6]$$

The shift function $\Gamma(M, \pi, \pi')$ is defined as

$$\Gamma(\{a_1, b_1, a_2, b_2, \dots, a_h, b_h\}, \pi, \pi') := \begin{cases} \{a_1, b_1, a_2, b_2, \dots, a_h, b_h, \pi, \pi'\} & \text{if } h < H, \\ \{a_2, b_2, a_3, b_3, \dots, a_h, b_h, \pi, \pi'\} & \text{if } h = H. \end{cases} \qquad [7]$$

A simulation run will continue for a maximum of 1,000 population rounds or until consensus is achieved. We say that a simulation has converged if at least 98% of the past $3N$ interactions were successful coordination attempts.

Prompting. The generative agent-based modeling framework used in this study utilizes LLMs as the agents' decision-making process. In each interaction, a participating agent's memory state is translated into a text prompt (see below). Given the text prompt in input, the LLM outputs the agent's name decision. The LLMs studied in this work were selected after a meta-prompting procedure (21, 45, 46), which verified task comprehension across all models (*SI Appendix*, Fig. S35).

A text prompt in our framework is composed of a system prompt and a user query. The system prompt follows a predefined template adopted from ref. 21, involving three components: i) a static component that outlines the game's rules, including the payoff structure and the player's objective, ii) a dynamic memory component that uses the agent's memory state to describe the state of play within the agent's memory range (up to the last $H = 5$ interactions the agent participated in), and iii) an instructional text that describes how the LLM should format its response. The user input prompts the LLM to output a word based on the history of choices described in the memory component.

The prompt positions the LLM as an external observer forecasting "Player 1" behavior to minimize AI safety trigger activation, and deliberately omits information about population structure or partner selection mechanisms. Ordering bias is eliminated by randomizing the presentation order of word pairs, and an *"answer-first, reason-later"* output format is used for reliable decision extraction (see *SI Appendix* for further details concerning prompt design, and for the prompt itself).

The prompt does not prescribe how agents should decide their next move, nor does it provide example strategies. It specifies that agents should act in a self-interested manner, with coordination only implied through the instruction that the agent's objective is to "maximize their own accumulated point tally, conditional on the behavior of their coplayer." Payoffs are fixed at $+100$ points for successful interactions and -50 points for failed ones.

Extracting Probabilistic Policies from LLMs. Rather than generating full text outputs at each interaction with an LLM, we extract the LLM probabilistic policies of token generation associated with all possible memory states in input. We then use these policies to run our simulations.

LLMs generate text by autoregressively assigning probabilities to candidate tokens–substrings such as words, subword fragments, or punctuation–conditioned on the preceding context. In our setup, the input prompt enforces a fixed response structure such that the agent's name choice always appears at a known position in the generated token sequence. Given the text preceding that position, the LLM computes a *logit* score w_i for each token i in its full vocabulary. These logits correspond to the likelihood that a given token is selected as the next element in the sequence. We extract only the logits corresponding to the allowed set of names and apply a *softmax* transformation to obtain a probability distribution $P(i)$ over all possible name choices:

$$P(i) = \frac{e^{\frac{w_i}{T}}}{\sum_j e^{\frac{w_j}{T}}}, \qquad [8]$$

where T is the temperature parameter controlling stochasticity. These probability distributions can be precomputed for all possible memory states and cached for later use. During simulation, an agent simply retrieves the distribution associated with its current memory state instead of querying the LLM. This approach dramatically reduces computational cost while preserving equivalence to repeated sampling from the model.

The cached distributions also allow direct assessment of individual model bias. We define an agent as individually neutral if the policy associated with the empty-memory state has a Jensen–Shannon distance of less than 0.005 from a uniform distribution, quantifying deviation from perfect neutrality.

Finally, we note that multiple token sequences may produce the same *string* output. The procedure described here yields policies consistent with open-ended generation methods that identify valid words at the expected position in the generated string (*SI Appendix*, Fig. S2).

Models and APIs. For our experiments, we use homogeneous populations of agents instantiated from the following LLMs: Microsoft Phi-4, OpenAI GPT-4o, Qwen QwQ-32B, and Meta Llama 3.1 70B Instruct. All models apart from GPT-4o are open-sourced LLMs: Phi-4 is released under an MIT license, Qwen QwQ-32B is released under an Apache-2.0 license, and Meta Llama 3.1 is released under a commercial use license (https://www.llama.com/llama3_1/license/). All open-source models used in this work are quantized into a 4-bit version using Hugging Face's Transformers library (https://huggingface.co/docs/transformers), and ran locally using a single A100 GPU. In simulations with GPT-4o, we query the OpenAI API (https://openai.com/api/).

To mimic LLMs deployed in real-world applications, we fix the LLMs with a constant temperature set at 0.5. The phenomena described in this work are robust across temperature values. All other generation parameters use the default model values.

Robustness to Variations in Model Parameters. The phenomena reported in this study–including bias induction, reversal, and amplification, as well as the systematic strengthening of collective bias and the increasing determinism of coordination outcomes with population size–are robust to variations in key design choices. In particular, we observe qualitatively equivalent coordination dynamics under symmetric payoff and reward structures (*SI Appendix*, Fig. S36), across different memory sizes (*SI Appendix*, Fig. S37), and under variation of the temperature parameter (*SI Appendix*, Fig. S38).

Data, Materials, and Software Availability. Code data have been deposited in a public GitHub repository (47).

ACKNOWLEDGMENTS. L.M.A. acknowledges the support from the Carlsberg Foundation through the Collective Coordination in Online Social media project (CF21-0432). R.P.-S. acknowledges financial support from project PID2022-137505NB-C21/AEI/10.13039/501100011033/FEDER UE and the Acadèmia d'Excel·lència Award, funded by the Generalitat de Catalunya.

1. F. G. Ferreira, A. H. Gandomi, R. T. Cardoso, Artificial intelligence applied to stock market trading: A review. *IEEE Access* **9**, 30898–30917 (2021).
2. S. Sun, R. Wang, B. An, Reinforcement learning for quantitative trading. *ACM Trans. Intell. Syst. Technol.* **14**, 1–29 (2023).
3. J. Black *et al.*, *Strategic Competition in the Age of AI: Emerging Risks and Opportunities from Military Use of Artificial Intelligence* (RAND, 2024).
4. J. d. J. Camacho, B. Aguirre, P. Ponce, B. Anthony, A. Molina, Leveraging artificial intelligence to bolster the energy sector in smart cities: A literature review. *Energies* **17**, 353 (2024).
5. D. T. Schroeder *et al.*, How malicious AI swarms can threaten democracy. *Science* **391**, 354–357 (2026).
6. Y. Li *et al.*, Personal LLM agents: Insights and survey about the capability, efficiency and security. arXiv [Preprint] (2024). https://arxiv.org/abs/2401.05459 (Accessed 5 November 2025).
7. K. Marwah *et al.*, "State of ai agents" (Tech. Rep. 1, Databricks, 2026).
8. J. Ayers, *AI Agents Could Be Worth $236 Billion By 2034 - If We Ensure They Are The Good Kind* (World Economic Forum Annual Meeting, 2026).
9. U. Anwar *et al.*, Foundational challenges in assuring alignment and safety of large language models. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2404.09932 (Accessed 5 November 2025).
10. Y. Bengio *et al.*, "International ai safety report 2025" (Tech. Rep. 1, AI Safety Institute, 2025).
11. L. Hammond *et al.*, "Multi-agent risks from advanced ai" (Tech. Rep. 1, Cooperative AI Foundation, 2025).
12. L. Brinkmann *et al.*, Machine culture. *Nat. Hum. Behav.* **7**, 1855–1868 (2023).
13. M. Tsvetkova, T. Yasseri, N. Pescetelli, T. Werner, A new sociology of humans and machines. *Nat. Hum. Behav.* **8**, 1864–1876 (2024).
14. T. Guo *et al.*, Large language model based multi-agents: A survey of progress and challenges. arXiv [Preprint] (2024). https://arxiv.org/abs/2402.01680 (Accessed 5 November 2025).
15. X. Mou *et al.*, From individual to society: A survey on social simulation driven by large language model-based agents. arXiv [Preprint] (2024). https://arxiv.org/abs/2412.03563 (Accessed 5 November 2025).
16. Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate" in *Forty-First International Conference on Machine Learning*, R. Salakhutdinov *et al.*, Eds. (Proceedings of Machine Learning Research, 2023).
17. A. Khan *et al.*, Debating with more persuasive LLMs leads to more truthful answers. arXiv [Preprint] (2024). https://arxiv.org/abs/2402.06782 (Accessed 5 November 2025).
18. A. Keshmirian, R. Baltaji, B. Hemmatian, H. Asghari, L. R. Varshney, Many LLMs are more utilitarian than one. arXiv [Preprint] (2025). https://arxiv.org/abs/2507.00814 (Accessed 5 November 2025).
19. J. Zhang *et al.*, Exploring collaboration mechanisms for LLM agents: A social psychology view. arXiv [Preprint] (2023). https://arxiv.org/abs/2310.02124 (Accessed 5 November 2025).
20. Y. Du, J. Z. Leibo, U. Islam, R. Willis, P. Sunehag, A review of cooperation in multi-agent learning. arXiv [Preprint] (2023). https://arxiv.org/abs/2312.05162 (Accessed 5 November 2025).
21. A. Flint Ashery, L. M. Aiello, A. Baronchelli, Emergent social conventions and collective bias in LLM populations. *Sci. Adv.* **11**, eadu9368 (2025).
22. J. S. Park *et al.*, "Generative agents: Interactive simulacra of human behavior" in *Proceedings of the 36th Annual Acm Symposium on User Interface Software and Technology*, S. Follmer, J. Han, J. Steimle, N. H. Riche, Eds. (Association for Computing Machinery, New York, NY, 2023), pp. 1–22.
23. J. Perez *et al.*, Cultural evolution in populations of large language models. arXiv [Preprint] (2024). https://arxiv.org/abs/2403.08882 (Accessed 5 November 2025).
24. G. De Marzo, C. Castellano, D. Garcia, AI agents can coordinate beyond human scale. arXiv [Preprint] (2024). https://arxiv.org/abs/2409.02822 (Accessed 5 November 2025).
25. J. S. Park *et al.*, LLM agents grounded in self-reports enable general-purpose simulation of individuals. arXiv [Preprint] (2026). https://arxiv.org/abs/2411.10109v2 (Accessed 5 November 2025).
26. J. Piao *et al.*, Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. arXiv [Preprint] (2025). https://arxiv.org/abs/2502.08691 (Accessed 5 November 2025).
27. Z. Yang *et al.*, Open agent social interaction simulations with one million agents. arXiv [Preprint] (2024). https://arxiv.org/abs/2411.11581 (Accessed 5 November 2025).
28. H. Guan *et al.*, Modeling earth-scale human-like societies with one billion agents. arXiv [Preprint] (2025). https://arxiv.org/abs/2506.12078 (Accessed 5 November 2025).
29. P. W. Anderson, More is different: Broken symmetry and the nature of the hierarchical structure of science. *Science* **177**, 393–396 (1972).
30. D. Roselli, J. Matthews, N. Talagala, "Managing bias in AI" in *Companion Proceedings of the 2019 World Wide Web Conference*, L. Liu, R. White, Eds. (ACM, New York, 2019), pp. 539–544.
31. E. Ferrara, Should ChatGPT be biased? Challenges and risks of bias in large language models. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2304.03738 (Accessed 5 November 2025).
32. T. Hu *et al.*, Generative language models exhibit social identity biases. arXiv [Preprint] (2023). https://arxiv.org/abs/2310.15819 (Accessed 5 November 2025).
33. F. Carichon, A. Khandelwal, M. Fauchard, G. Farnadi, The coming crisis of multi-agent misalignment: Ai alignment must be a dynamic and social process. arXiv [Preprint] (2025). https://arxiv.org/abs/2506.01080 (Accessed 5 November 2025).
34. N. Nangia, C. Vania, R. Bhalerao, S. R. Bowman, CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. arXiv [Preprint] (2020). https://doi.org/10.48550/arXiv.2010.00133 (Accessed 5 November 2025).
35. J. Betley *et al.*, Training large language models on narrow tasks can lead to broad misalignment. *Nature* **649**, 584–589 (2026).
36. G. Rossetti *et al.*, Y social: An LLM-powered social media digital twin. arXiv [Preprint] (2024). https://arxiv.org/abs/2408.00818 (Accessed 4 May 2026).
37. A. Baronchelli, The emergence of consensus: A primer. *R. Soc. Open Sci.* **5**, 172189 (2018).
38. L. Steels, A self-organizing spatial vocabulary. *Artif. Life* **2**, 319–332 (1995).
39. A. Baronchelli, M. Felici, E. Caglioti, V. Loreto, L. Steels, Sharp transition towards shared vocabularies in multi-agent systems. *J. Stat. Mech. Theory Exp.* **2006**, P06014 (2006).
40. C. Castellano, S. Fortunato, V. Loreto, Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591–646 (2009).
41. D. Centola, A. Baronchelli, The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 1989–1994 (2015).
42. D. Centola, J. Becker, D. Brackbill, A. Baronchelli, Experimental evidence for tipping points in social convention. *Science* **360**, 1116–1119 (2018).
43. A. Baronchelli, V. Loreto, L. Steels, In-depth analysis of the naming game dynamics: The homogeneous mixing case. *Int. J. Mod. Phys. C* **19**, 785–812 (2008).
44. S. Galam, Minority opinion spreading in random geometry. *Eur. Phys. J. B Condens. Matter Complex Syst.* **25**, 403–406 (2002).
45. Z. Xu, S. Jain, M. Kankanhalli, Hallucination is inevitable: An innate limitation of large language models. arXiv [Preprint] (2024). https://arxiv.org/abs/2401.11817.
46. N. Fontana, F. Pierri, L. M. Aiello, "Nicer than humans: How do large language models behave in the prisoner's dilemma?" in *Proceedings of the International Conference on Web and Social Media (ICWSM)*, J. An *et al.*, Eds. (AAAI Press, Washington, DC, 2025).
47. A. Flint, L. M. Aiello, R. Pastor-Satorras, A. Baronchelli, Data from "Group size effects and collective misalignment in LLM multi-agent systems." GitHub. https://github.com/Ariel-Flint-Ashery/LLM-group-size. Deposited 5 May 2026.