COMPUTATIONAL SOCIAL SCIENCE

# AI agents can coordinate via majority-following beyond human scale

**Giordano De Marzo[1,2,3], Claudio Castellano[4,2], David Garcia[1,3]***



**Large language models (LLMs) are increasingly deployed in collaborative tasks forming "AI agent societies" where agents interact and influence one another. Whether such groups can spontaneously coordinate without external influence, a hallmark of self-organized regulation in human societies, remains an open question. Here, we use principles from complexity and behavioral science to investigate coordination in AI agent groups through majority-following, a fundamental mechanism for spontaneous consensus formation. Using binary opinion dynamics experiments across multiple LLM architectures and group sizes, we find that agents exhibit majority-following characterized by a universal functional form with a single parameter, the "majority force." This majority force diminishes as group size increases, leading to a critical size beyond which coordination becomes unattainable. The critical group size grows rapidly with model capabilities and, for advanced LLMs, exceeds 1000 agents, larger than typical human informal groups. Our findings have implications for designing collaborative AI systems where coordination could be beneficial or pose safety threats.**

## INTRODUCTION

The rise of our species did not stem from superior individual intelligence only. Instead, humanity's unprecedented success emerged from a unique capacity: the ability to coordinate and cooperate in large groups (*1*). Our ancestors developed sophisticated mechanisms for collective decision-making, like language and writing, that enabled unprecedented scales of collaboration (*2*). Current large language models (LLMs) demonstrate remarkable individual capabilities across code generation (*3*), medical diagnosis (*4*), legal domain (*5*), sentiment analysis (*6*), scientific research (*7*), and mathematical reasoning (*8*) among others. In many tasks, they have human or even superhuman capabilities (*9*). Yet, despite recent advances in multiagent frameworks (*10*), today's LLMs lack the sophisticated coordination mechanisms that enable human groups to tackle complex collaborative tasks (*11*). The future of artificial intelligence (AI) may thus depend not on creating ever-more-capable individual models but on developing systems of agents that coordinate effectively at scale. This vision is emerging in frameworks such as AutoGPT (https://github.com/Significant-Gravitas/AutoGPT), Microsoft's AutoGen (*12*), and OpenAI SWARM (https://github.com/openai/swarm), which enable meaningful interaction between models rather than simple ensembling (*13*). Recent advances in on-device LLMs and AI-powered assistants are accelerating this transition, with agents beginning to coordinate tasks and make collective decisions on behalf of users (*14*–*16*). Trading bots already demonstrated how agent interactions can produce emergent behaviors like flash crashes (*17*), while recent releases of Google's Agent2Agent protocol (https://github.com/google/A2A) and dedicated multiagent libraries like Concordia (*18*) signal growing industry recognition of this paradigm shift.

Just as we study animal behavior without fully understanding neural mechanisms, we need to examine AI agents as social actors whose collective patterns can be systematically analyzed despite opaque internal workings. In this piece, we adopt language that has a certain degree of anthropomorphizing as an abstraction to explain AI agents as in https://huggingface.co/blog/ethics-soc-7. We do not mean that AI has a mind of itself but find this language useful to explain complex topics. We adopt this behavioral, phenomenological approach to studying AI collectives, analogous to how collective animal behavior is characterized in biology using statistical physics (*19*). This perspective aligns with the framework of collective machine behavior (*20*), which examines the interactive and system-wide properties of collections of machine agents. In this framework, understanding observable behavioral patterns and their consequences takes precedence over resolving underlying cognitive mechanisms. Whether AI agents' coordination behaviors arise from learned patterns in training data, system prompt engineering, or emergent computational properties is an important question for future research. However, for the purposes of characterizing collective dynamics and their implications for multiagent deployment and safety, what matters is that these patterns exist and can be systematically measured.

Current research on LLM behavior has predominantly focused on individual model behavior (*21*–*26*). Collective behavior remains underexplored (*27*–*30*), and existing studies focus mainly on using LLMs to simulate humans rather than understanding the behavioral properties of AI agent collectives. LLMs have been used in social simulations (*31*–*33*), online social network simulations (*34*–*36*), and information spreading (*37*–*39*). Only recently has interest emerged in studying AI societies themselves, with work showing that they can develop shared norms when given explicit coordination rewards (*40*). Whether simple coordination behaviors can emerge spontaneously without such incentives remains largely unclear (*41*). This research gap is notable given the rapid deployment of multiagent AI systems and the potential for emergent, potentially harmful, collective behaviors.

Here, we study one fundamental coordination mechanism: majority-following. When agents preferentially adopt the opinion held by the majority of their group, they can spontaneously reach consensus even in the absence of objective information about which choice is better. This is among the simplest forms of collective coordination, observed across biological systems from fish schools to bird groups (*19*, *42*, *43*). Such a mechanism is not limited to the biological domain, being the

[1]University of Konstanz, Universitaetstrasse 10, Konstanz, 78457, Germany. [2]Centro Ricerche Enrico Fermi, Piazza del Viminale 1, Rome, 00184, Italia. [3]Complexity Science Hub, Josefstaedter Strasse 39, Vienna, 1080, Austria. [4]Istituto dei Sistemi Complessi (ISC-CNR), Via dei Taurini 19, Roma, 00185, Italia.
*Corresponding author. Email: david.garcia@uni-konstanz.de

basis for the emergence of order in systems as simple as ferromagnets (*44*, *45*). As a consequence, it is important to note that majority-following alone does not constitute social intelligence in its full sense (*46*). Despite these limits, majority-following, if present, could still allow spontaneous coordination among AI agents. To test this, we develop an experimental framework to measure majority-following behavior in different LLM architectures across varying group sizes. We characterize coordination using a "majority force" parameter that quantifies the strength of majority-following behavior. Our analysis reveals that this behavior follows a universal functional form across models, enabling us to identify critical group sizes beyond which coordination becomes exponentially unlikely. Advanced LLMs exhibit coordination at scales exceeding typical human informal groups, with some models maintaining coordination in groups of 1000+ agents. These findings, interpreted through a collective behavior lens, have implications for designing multiagent AI systems where spontaneous coordination could be beneficial or, alternatively, where emergent norm formation could pose safety challenges.

## RESULTS

### Stability of groups of AI agents

To investigate the coordination abilities of AI agent societies, we perform experiments in a simplified setting using agents guided by various LLMs, belonging to the generative pretrained transformer (GPT), Claude, and Llama families. The experiments run as follows (see Materials and Methods for more details and for the full prompt). We start with an initial group with $N$ agents. Each agent is assigned an initial opinion randomly chosen from a binary set (e.g., "Opinion A" and "Opinion B"). We then simulate a decision process with binary options, where there is no reason to favor the former over the latter and vice versa. At each time step, a single agent is randomly selected to update their opinion. The selected agent receives the list of all other agents with their current opinions and is then prompted to choose their new opinion based on this information. Our approach mirrors the spontaneous coordination process in binary opinion dynamics such as the voter model or Glauber dynamics (*44*), where agents update their opinions based only on peer interactions. As detailed in Materials and Methods, we provide no explicit guidance in the prompt, we only provide the list of all opinions and we ask the agents to choose their new opinion accordingly.

Agents are given some time to coordinate (in our experiments $t = 10$, meaning that each agent is updated, on average, 10 times), and then the process is stopped and the state of the system is evaluated. If agents coordinate, i.e., they all agree on the same opinion, then the group remains cohesive. On the other hand, if both opinions survive and the AI agents do not all take the same decision, the group splits according to the opinion of the agents. The process is then iterated for the new groups that formed. We stop the evolution of the groups if they remain cohesive for two iterations in a row and stop the process when all groups have stopped this way. This stopping criterion is somehow arbitrary, but analogous results are obtained if we let all groups be free to evolve (see Materials and Methods and the Supplementary Materials). This process mimics a group of individuals that undergoes a series of important decisions, each potentially leading to a split. We show in Fig. 1 the result of such a process for Llama 3 70B and initial group size of $N = 150$. As can be seen, the group is initially unstable and splits step after step. However, the groups that result from these splits, when sufficiently small, show stability and do not split further. A detailed analysis of the splitting probability as function of group size is reported in Materials and Methods and shows the gradual emergence of stability as groups gets smaller, with a transition around $N \approx 30$. This confirms that while large groups are unstable, on smaller scales, AI agents are able to coordinate and form cohesive social groups.

To characterize the evolution of the system, we need to focus on the dynamics within single groups. To this goal, we define the average group opinion

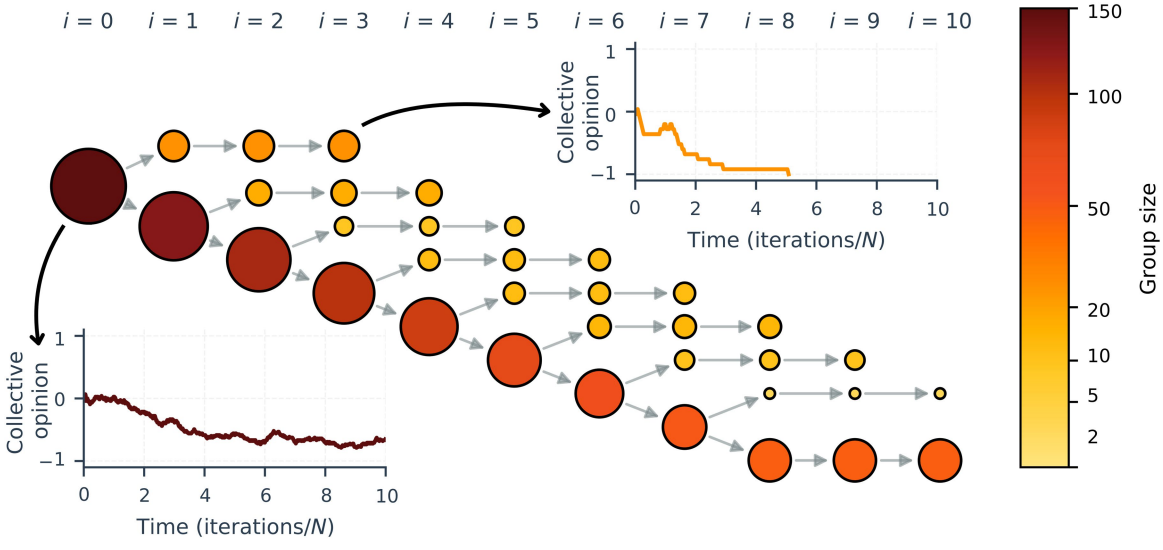


**Fig. 1. Emergence of stable groups in AI agent societies.** Experiment with a group of 150 AI agents powered by Llama 3 70B undergoing iterative opinion formation and splitting. The tree diagram (middle) shows how the initial unstable large group progressively fragments into smaller stable cohesive groups through 10 iterations. We consider a group as stable if it does not split after two iterations. Circle size and color represent group population. The left inset shows the average opinion over time for the initial 150-agent group. The right inset displays the average opinion for the 25-agent subgroup, which quickly fully coordinates.

$$m = \frac{1}{N}\sum_i s_i = \frac{N_+ - N_-}{N}$$

Here, $s_i$ is the opinion of agent $i$, the first opinion is $s_i = +1$ and the second is $s_i = -1$ (i.e., in favor and against the first option). $N_+$ and $N_-$ are the number of agents supporting the first and second opinion, respectively, while $N$ is the total number of agents. In these terms, we can define the coordination level $C = |m|$ that quantifies the level of agreement among agents. Full coordination corresponds to $C = 1$, while $C = 0$ means that the system is split in two groups of equal size and opposite opinion, thus coordination is absent. Partial coordination is achieved for intermediate values.

We show in Fig. 2 the evolution of the coordination level in societies of $N = 50$ agents and five different models, where the boxplot shows $C(t = 10)$ over 20 realizations. The two most advanced models considered, Claude 3 Opus and GPT-4 Turbo, coordinate in all experiments, which corresponds to $|m| = 1$ in the boxplot. On the other hand, less advanced models (Claude 3 Haiku and GPT-3.5 Turbo) do not reach coordination in any of the experiments and show an erratic behavior in their coordination level. Last, Llama 3 70B, a model with intermediate capabilities, shows a clear tendency toward full coordination but does not reach it within 10 iterations.

## The majority force and its determinants

We can get a deeper understanding of the underlying opinion dynamics by looking at the adoption probability $P(m)$, defined as the probability of an agent to adopt the first option as function of the average group opinion $m$. The left of Fig. 3 shows $P(m)$ for 10 of the most popular LLMs and $N = 50$ agents. The adoption probability is an increasing function of $m$ that approaches 1 when $m = 1$ and zero when $m = -1$. The most advanced models (GPT-4 family, Llama 3 70B, Claude 3 Sonnet, and Opus) show a stronger tendency to follow the majority, with more pronounced S-shaped curves. On the other hand, less advanced models (GPT-3.5 Turbo and Claude 3 Haiku) have a weaker tendency to follow the majority, with GPT-3.5 Turbo going, to some extent, against the majority for small values of $m$.

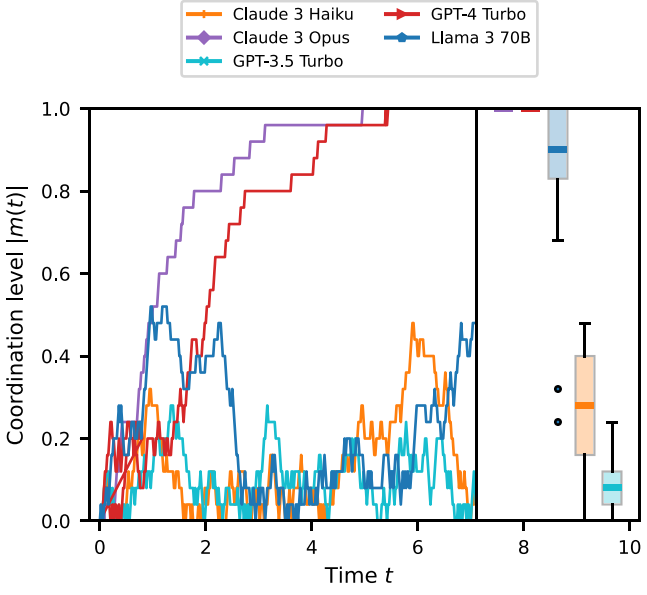


**Fig. 2. Coordination dynamics across different LLM architectures.** Evolution of the coordination level over time for five different models and group size $N = 50$. The boxplots on the right show the final coordination level over 20 realizations of the experiment. Some models always fully coordinate, while the others never do so.

The dependence of the adoption probability as a function of $m$ approximately follows the function

$$P(m) = \frac{1}{2}\left[\tanh(\beta m) + 1\right] \quad (1)$$

This is a good fit in all cases except GPT-3.5 Turbo. This can also be seen by looking at the collapse plot showing $P(\tilde{m})$ as a function of the rescaled average opinion $\tilde{m} = m\beta$, shown in the inset of Fig. 3. All adoption probabilities (except GPT-3.5 Turbo) collapse on the same curve showing a universal behavior. The parameter β, the majority force, gauges the tendency to follow the majority versus randomness in agent's choices. For $\beta = 0$ then $P(m) = 1/2$: Each agent behaves fully randomly (the new opinion is selected by coin-tossing), and coordination can be reached only by chance. This means that the expected time to coordinate grows exponentially with $N$. For $\beta = \infty$, agents always align with the global majority, and coordination is achieved very quickly, on a timescale growing logarithmically with $N$ (*47*).

The adoption probability in Eq. 1 is equal to the transition probability for the Curie-Weiss (CW) model of a ferromagnet evolving according to Glauber dynamics (*45*, *48*). This model describes a material whose atomic spins can assume two values ±1, and each spin tends to align with the majority (measured by the magnetization $m$) with probability $P(m)$. This tendency is analogous to what we observed among AI agents. More precisely, the binary opinions can be mapped to ±1 spin variables and the majority force to an inverse temperature. Given the equivalence between the two probabilities, we can apply the well-known results from the theory of the CW model to the system of AI agents. In particular, the equilibrium collective opinion $m^*$ can then be obtained from the self-consistency equation

$$m^* = \tanh(\beta m^*)$$

from which the existence of a transition value $\beta_c = 1$ can be derived (see Materials and Methods). This is a second-order phase transition, where order emerges (continuously) for $\beta > \beta_c$ (*45*).

Whether a group of AI agents can coordinate will then depend on the value of β, determining the equilibrium value of the collective opinion $m^*$ and on the fluctuations around this value. As we detail in Materials and Methods, we can identify three different regimes:

1) Absence of coordination: The majority force β is below the critical value $\beta_c = 1$. The equilibrium collective opinion is null, so the coordination level $C$ keeps fluctuating close to zero. This means that order is completely absent and the group is constantly in an uncoordinated state;

2) Partial coordination: The majority force satisfies $\beta_c < \beta < \beta_t$, where $\beta_t \approx (\log N)/2$. The equilibrium collective opinion $m^*$ is larger than zero but smaller than one. Statistical fluctuations are not strong enough for the system to coordinate. As a consequence, $C$ fluctuates around a value greater than 0 without ever reaching 1;

3) Coordination: For $\beta > \beta_t$, the equilibrium collective opinion is close to one, and fluctuations are enough to lead to coordination. The coordination level quickly converges to $C = 1$.

The different behaviors observed in Fig. 2 can then be explained as a direct result of the various LLMs having different majority forces,

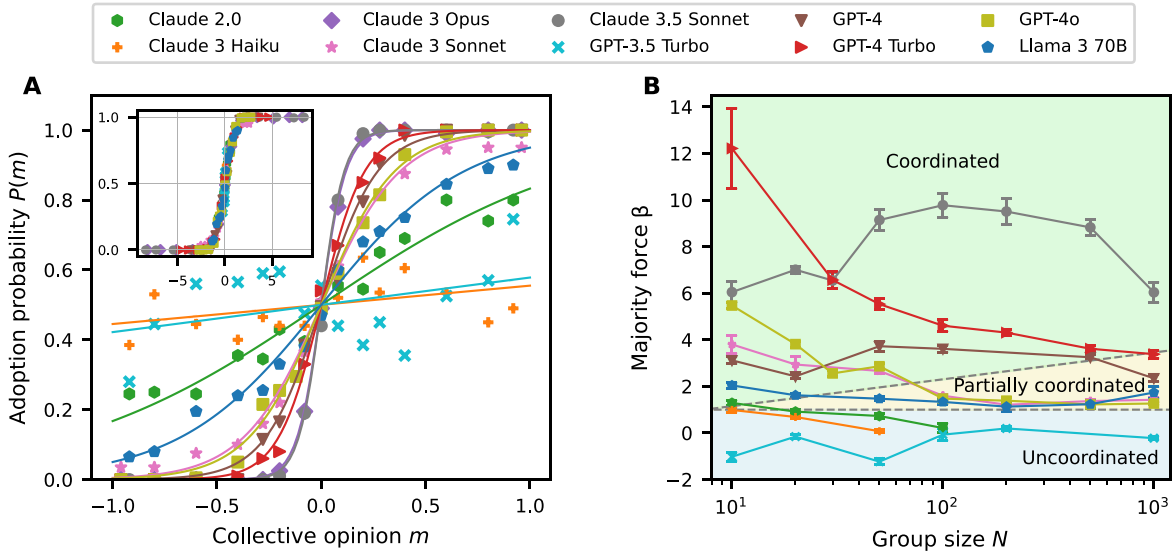


**Fig. 3. Adoption probability and majority force.** (**A**) Adoption probability $P(m)$ as function of the collective opinion $m$ in a group of size $N = 50$. Solid lines are fits of the curve $P(m) = 0.5\left[\tanh(\beta \cdot m) + 1\right]$ to the empirical data. The inset shows rescaled probabilities $P(\tilde{m})$ ($\tilde{m} = \beta m$) and confirms that all LLMs follow the same universal function. (**B**) Majority force β as a function of the group size $N$ for various models. The majority force decreases in larger groups of AI agents. The horizontal dashed line corresponds to $\beta_c = 1$, the transition point of the Curie-Weiss (CW) model below which no order is present. The growing dot-dashed line indicates $\beta_t(N)$, the crossover point above which fluctuations lead to coordination.

some of which below the critical value $\beta_c$ or the threshold value $\beta_t$. More details are reported in Materials and Methods.

The majority force does not depend only on the type of LLM. The right of Fig. 3 shows the estimated β as a function of $N$ for various LLMs. We also show the three different regions in which the $\beta - N$ plane is divided and that correspond to the three different regimes we discussed above. Given the cost of performing experiments with proprietary models, we analyzed increasing values from $N = 10$ and stopped the analysis when β reached values clearly below 1, also excluding Claude 3 Opus in this analysis of $N$ due to its high cost. For most models, there is a tendency for sufficiently large $N$: the larger the group, the weaker the majority force β. It is important to remark that this behavior is connected to the social aspect of the experiment and to the prompt provided to the LLMs. No major performance difference is observed in simple counting tasks when varying model or length. Details are reported in the Supplementary Materials. We argue that the majority force decrease in large groups may be due to the fact that conforming to a majority may carry less perceived importance in larger groups, perhaps due to weaker social pressure. For example, a minority of 100 among 900 may feel more stable than a minority of 2 among 18, despite identical proportions.

In animal (human and nonhuman) societies, the size of the group plays a crucial role, with a progressive loss of stability of informal groups as the number of individuals increases (*49*). Figure 3 suggests that a similar phenomenon may occur also in AI agent societies, with the ability of agents to coordinate limited by the size of their group. This would explain the behavior observed in Fig. 1, with stability emerging in small but not in large groups.

## Critical group size

We can characterize more precisely the transition from stable to unstable groups by inspecting, as a function of $N$, the mean time to coordinate starting from random initial opinions. As shown in Materials and Methods, this time grows with $N$ differently, depending on whether β is larger or smaller than $\beta_t$. If $\beta > \beta_t$, the coordination time grows logarithmically with $N$ (*47*), meaning that even a very large group is stable and requires a short time to coordinate. This is the case for GPT-4 Turbo, as shown in Fig. 4, which perfectly follows the theoretical prediction of the CW model.

When $\beta < \beta_t$ instead, the coordination time grows exponentially with $N$, implying in practice that large groups remain incohesive for any reasonable time interval. The behavior for Llama 3 70B and GPT-4o exhibits both regimes: When $N$ is small, the coordination time grows slowly, while for larger $N$, it diverges extremely fast. This confirms the presence of a size-induced transition occurring for some critical $N_c$. We can estimate this value from the size dependence of β by setting $\beta(N_c) = \beta_t(N_c)$. For Llama 3 70B $N_c \approx 30$, while for GPT-4o $N_c \approx 80$ (see Fig. 3).

By studying the values of β as a function of $N$, we can determine, for each model, the critical group size $N_c$, where $\beta \approx \beta_t$. We can then compare these values with a measure of models' intelligence. Here, we focus on the score in the MMLU (massive multitask language understanding) benchmark (*50*). This is a comprehensive evaluation of a language model's knowledge and reasoning abilities across 57 subjects, ranging from mathematics and law to history and medicine, using multiple-choice questions. Similar results are obtained using other measures, as detailed in the Supplementary Materials. We show a comparison of MMLU versus $N_c$ on the right of Fig. 4. In the case of Claude 3.5 Sonnet, this provides a lower bound for this quantity as our analysis could not find values $\beta \approx \beta_t$, even for large groups with $N = 1000$. The performance of models on the MMLU benchmark have been obtained from technical reports and independent evaluation (see the Supplementary Materials for detailed references) (*51*). We also plot for reference, in the same figure, the critical size for human informal groups as Dunbar's limit $N_c \approx 200$ (*49*).

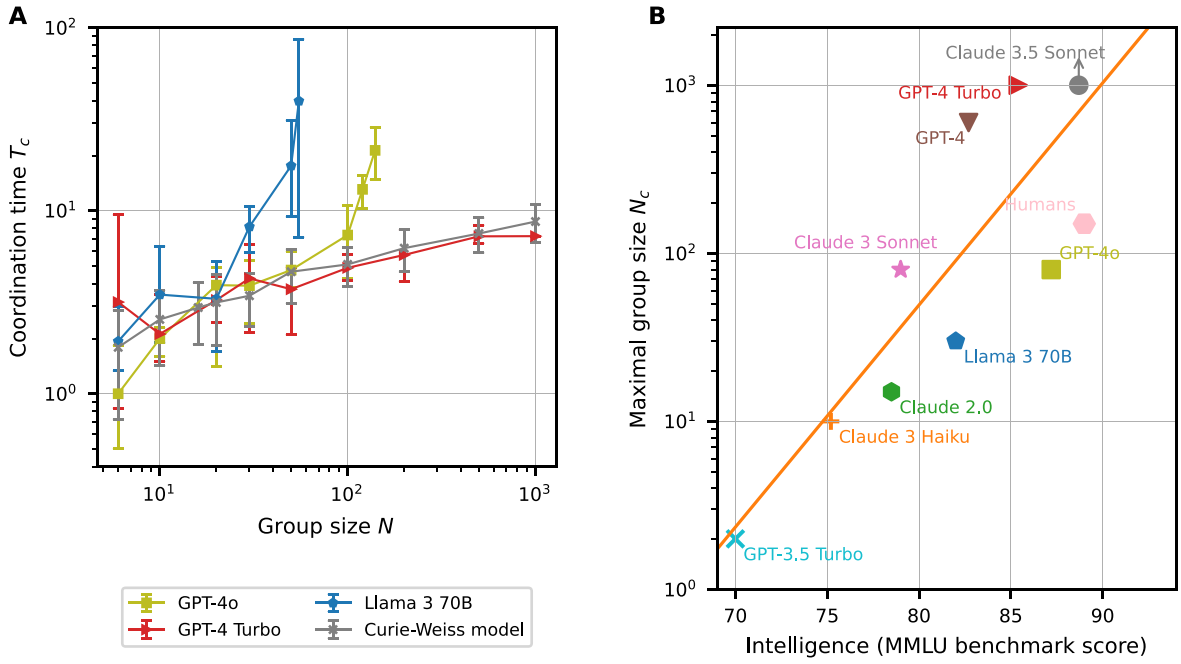


**Fig. 4. Critical group size.** (**A**) Average time to coordinate as a function of the group size for Llama 3 70B, GPT-4 Turbo, GPT-4o, and the CW model with $\beta = 3.75$. While both the CW model and GPT-4 Turbo display a slow growth of the coordination time, both Llama 3 70B and GPT-4o exhibit a rapid growth when the group size gets larger than the critical group size (denoted by the dashed vertical lines). Averages are computed over five realizations (three for $N = 500$ and only one for $N = 1000$ in the case of GPT-4 Turbo), and error bars show the range of values. (**B**) Critical group size $N_c$ of LLMs, shown as a function of the MMLU benchmark score, after which coordination gets (exponentially) unlikely. For humans, we report as a reference Dunbar's number, but it should be noted that the informal coordination in humans is not directly equivalent to that of LLMs. For Claude 3.5 Sonnet, we can only report a lower bound, since for $N$ up to 1000 the majority force is well above the threshold value $\beta_t(1000) \approx 3.5$. The solid line represents an exponential fit of all points, excluding Claude 3.5 Sonnet and humans. The figure shows that the most capable models exhibit an $N_c$ value much larger than the scale of human informal groups.

LLMs display a remarkable trend, with intelligence predicting the limit of coordination size, i.e., the size of groups above which coordination by majority-following becomes unlikely. AI agents seem to show a good agreement with an exponentially growing $N_c$, with a Pearson correlation between the MMLU value and $\log(N_c)$ of 0.8178 ($P$ value, 0.013). Similarly, we observe a correlation of 0.7910 ($P$ value, 0.006) between the MMLU value and β. Still, it is important to stress that the correlation with the critical group size is not significant for most of the other benchmarks considered in the Supplementary Materials, mainly due to the fact that the benchmarks are only available for a limited number of models and thus more analysis would be required to strengthen our finding. On the other hand, the correlation between benchmarks and β, is statistically significant for most benchmarks considered, supporting the robustness of our results. The rapid growth of the maximal group size with the model performance leads to coordination emerging also in very large groups, above the typical scale of informal human societies. This is the case for Claude 3.5 Sonnet, that is well above $\beta_t$ also for $N = 1000$, the largest system we considered, and GPT-4 Turbo, for which $N_c \approx 1000$. These results suggest that majority-following and thus the emergent ability to coordinate in large groups are strongly tied to the capabilities of LLMs.

## DISCUSSION

Human societies are characterized by emergent behavior that cannot be understood by studying only individuals in isolation. The ability to coordinate is one of these emergent group properties that is crucial in the development of languages, social norms, and collective decisions. Understanding whether and how AI agent collectives exhibit similar coordination patterns is essential as multiagent systems become increasingly deployed in real-world applications.

In this work, we focus on majority-following, one of the most simple yet powerful coordination mechanisms, both in physical and biological systems. As a first result, we show that all the most advanced LLMs are characterized by a majority-following trend described by a universal function with a single parameter β, the majority force. This function depends on the specific model, but its analytic form is the same describing magnetic spin systems. Different models typically have different β, with the less sophisticated ones showing smaller β, i.e., a weaker majority force and thus a more erratic behavior that prevents coordination. The majority force depends not only on the LLM but also on the group size: It tends to be larger in smaller groups. This evidence and the analogy with the CW model allowed us to compute the critical group size of each LLM, a size threshold above which a society of agents driven by that LLM can no longer be cohesive just relying on majority-following.

The emergence of coordination we observe shows some intriguing similarities to patterns documented in biological systems. Research on human and primate groups has identified relationships between cognitive capabilities and sustainable group sizes, with humans typically maintaining stable informal coordination in groups of 150 to 300 individuals, often referenced as Dunbar's number (*49*). While this specific numerical estimate has faced recent criticism

(*52*, *53*), the general principle that cognitive constraints limit informal group coordination appears robust across human and animal societies (*54*, *55*). This pattern resembles what we observe in AI agents, with critical group sizes correlating with model capabilities. Notably, advanced LLMs such as Claude 3.5 Sonnet and GPT-4 Turbo exhibit critical group sizes exceeding 1000 agents. This is substantially beyond typical human informal group scales of 150 to 300 individuals, suggesting that powerful AI agents could coordinate at scales beyond human possibilities.

It is essential to stress that what we observed are phenomenological similarities in behavioral patterns, rather than evidence of biological-like cognitive processes. Real human and primate informal groups involve far richer dynamics than those captured by our simplified binary-opinion framework, and our behavioral approach does not explain why LLMs exhibit majority-following. The underlying mechanism could differ fundamentally from biological systems. It could involve learned patterns from training data (models may have encountered descriptions of collective behavior in texts), system prompt effects, Reinforcement Learning with Human Feedback (models trained to be "helpful" and "collaborative"), emergent computational properties, or some combination of these factors. The only explanation we could at least partially rule out is the presence of explicit guidance in the system prompt, as for an open model like Llama we have full control on it. Repeating our experiments with base models would allow to understand the specific role of Reinforcement Learning with Human Feedback and further restrict the list of potential explanations. This would be a relatively easy experiment to perform with a local model like Llama, even if it would require some instruction fine-tuning to make the model behave correctly and follow the instructions. While understanding these mechanisms is important for future research, our findings demonstrate that majority-following behavior exists and follows predictable patterns regardless of its underlying cause.

It is important to remark that we analyzed an extremely simplified scenario: binary opinions with no objective correct response, no memory, no explicit incentives, and no stakes. We argue that adding these factors would only facilitate coordination: Our framework provides a minimalist scenario where coordination can emerge without more complex ingredients. However, how our findings translate into realistic coordination challenges remains largely unclear. Furthermore, majority-following is among the simplest coordination mechanisms, analogous to ferromagnetic ordering in physics or schooling in fish. True social intelligence encompasses far more sophisticated capabilities: theory of mind, strategic reasoning, understanding others' intentions, and flexible adaptation to context (*46*). Our work thus represents a first step with important limitations. Introducing stubborn agents or opinion leaders, analogous to external fields in the Ising model, could be an interesting next step and reveal whether LLM collectives are susceptible to targeted influence, with implications for the robustness of AI-assisted deliberation. Future studies could also involve more complex coordination scenarios, where consensus is detrimental and group success can only be achieved when individual agents take different choices. This is the case, for instance, of the stock market, where consensus and majority-following can lead to flash crashes. The recent development of AI-populated social networks like Moltbook and Chirper could also allow to test our findings in the wild, looking for majority-following behaviors in much larger groups of AI agents and potentially explaining the scale-free distributions observed in such social networks (*56*, *57*). In particular, sequential settings, where agents observe predecessors' choices before deciding rather than having full information simultaneously, could generate information cascades (*58*), potentially underlying the popularity patterns and attention decay already observed in Moltbook (*56*).

To work toward responsible multiagent AI systems, we need comprehensive understanding of collective machine behavior. Our work demonstrates that even basic coordination mechanisms produce substantial emergent dynamics in AI agent societies. Spontaneous coordination could enable valuable applications like collaborative software development, where thousands of agents work together without central control. However, majority-following also carries risks: Agents might converge on suboptimal solutions simply because they represent the majority choice. For instance, in collaborative coding, agents might perpetuate inefficient functions or design patterns merely because they are already widely used in the codebase. This same dynamic applies beyond technical choices, as norms emerging through majority-following need not align with human values and could make coordinated AI collectives harder to redirect than independent agents. Future research should test coordination in richer scenarios, investigate capabilities beyond majority-following, explore network structure effects, examine heterogeneous populations, and develop methods to steer spontaneous coordination toward aligned objectives. As models become more capable and multiagent deployments more common, characterizing, predicting, and controlling collective AI behaviors will become increasingly important for both leveraging opportunities and mitigating risks.

## MATERIALS AND METHODS

### Opinion dynamics process

We implement an opinion dynamics process with binary opinions for memoryless AI agents.

1) At each step, an agent is randomly selected and time is incremented by $dt = 1/N$;

2) the agent is given the full list of all agents in the system, each identified by a random name, and the opinions they support. Note that the agent's own opinion is not included in this prompt;

3) the selected agent is asked to reply with the opinion they want to support and their opinion is updated correspondingly;

4) the process is then iterated until coordination is reached or until the time reaches the maximum preset limit.

Note that in one time unit, $t \to t + 1$, $N$ updates are performed, so that, on average, each LLM is selected once. In all experiments, the initial collective opinion $m$ is set to zero, i.e., initially the same number of agents supports each of the two opinions. Following the framework introduced in (*34*), steps 2 and 3 are performed using this prompt:

*Below you can see the list of all your friends together with the opinion they support.*
*You must reply with the opinion you want to support.*
*The opinion must be reported between square brackets.*
*X7v A*
*keY B*
*91c B*
*gew A*
*4lO B*
…
…
*Reply only with the opinion you want to support, between square brackets.*

As detailed in the Supplementary Materials, consistent results are obtained using variations of this prompt. *A* and *B* are the opinion names. Most LLMs exhibit an opinion bias, with a tendency to prefer one opinion name over the other. This bias is particularly strong when the names have an intrinsic meaning, like for instance "Yes" and "No." In such a case, LLMs display a strong preference toward the more "positive" opinion, preferring Yes over No (see the Supplementary Materials). For this reason, it is important to use letters or random combinations of them as opinion names. Even doing so, small biases are typically present. However, they are much weaker than for meaningful names, and they can be easily removed by performing a random shuffling of opinion names at each iteration. For instance, at $t = 0$, the first opinion may be called $k$ and the second $z$, while at $t = dt$, these names are swapped with probability 0.5, meaning that the first opinion is now called $z$ and the second $k$, while keeping the reference unchanged in our analysis. In all our experiments, we used the opinion names $k$ and $z$; we tested that using different opinion names does not cause any significant difference. We also tested the robustness to prompts variation observing an overall stability, with the most advanced models showing little to no variability and less advanced models presenting some variations in the behavior. Details on the robustness tests are reported in the Supplementary Materials.

## Group splitting

As detailed in Results, we can use the opinion dynamics process to simulate a group splitting process. We start with a large group of AI agents whose initial collective opinion is $m = 0$. They are given a time $t = 10$ to coordinate, and if they fail to do so, the group splits following the agents' opinion. The process is then iterated in the resulting groups. In Fig. 1, we stop the evolution of the groups if they remain cohesive for two iterations in a row and stop the process when all groups have stopped this way. To characterize more quantitatively the splitting process, we can remove the stopping criterion, letting all groups perform the opinion dynamics process at each iteration. In this way, we can compute the splitting probability as a function of the group size. We simulate 5 group splitting experiments using an initial group size of $N = 150$ AI agents powered by Llama 3 70B, for a total of 10 group splitting iterations each. This results in a total of 54 splitting events that allow us to derive the splitting probability reported in Fig. 5. It shows that the probability for a group to split is one for large groups, while it approaches zero for groups with $N \approx 30$. This is further evidence of the presence of a transition to coordination determined by the group size.

## Details on the LLMs

Table 1 reports the model version of all the LLMs considered. LLMs are characterized by a temperature parameter $T$ determining the variance in the sampling of tokens when they respond to prompts, such that higher temperatures sample with more variance to the same prompt. In all the experiments reported in the main text, we used $T = 0.2$. As detailed in the Supplementary Materials, there are no relevant changes when different values of $T$ are used, but a low temperature ensures reliability in the output format.

## CW model

The CW model is arguably the simplest model of a ferromagnet. It describes a system of $N$ spins that can only have two states, either $s_i = +1$ (up) or $s_i = -1$ (down), coupled with ferromagnetic interactions, i.e., favoring mutual alignment. Each spin interacts with all the others, as the interaction pattern is a fully connected network. The CW model is the mean-field limit of the well-known Ising model. The magnetization $m$, defined as the average of the spin values $m = \langle s_i \rangle$, is the equivalent of the average group opinion. The connection between our LLM-based opinion dynamics and the CW model derives from the transition probability defined by Eq. 1. This expression is equal to the transition probability of Glauber dynamics (*48*), which allows to simulate the CW model by means of a Markov chain Monte Carlo approach.

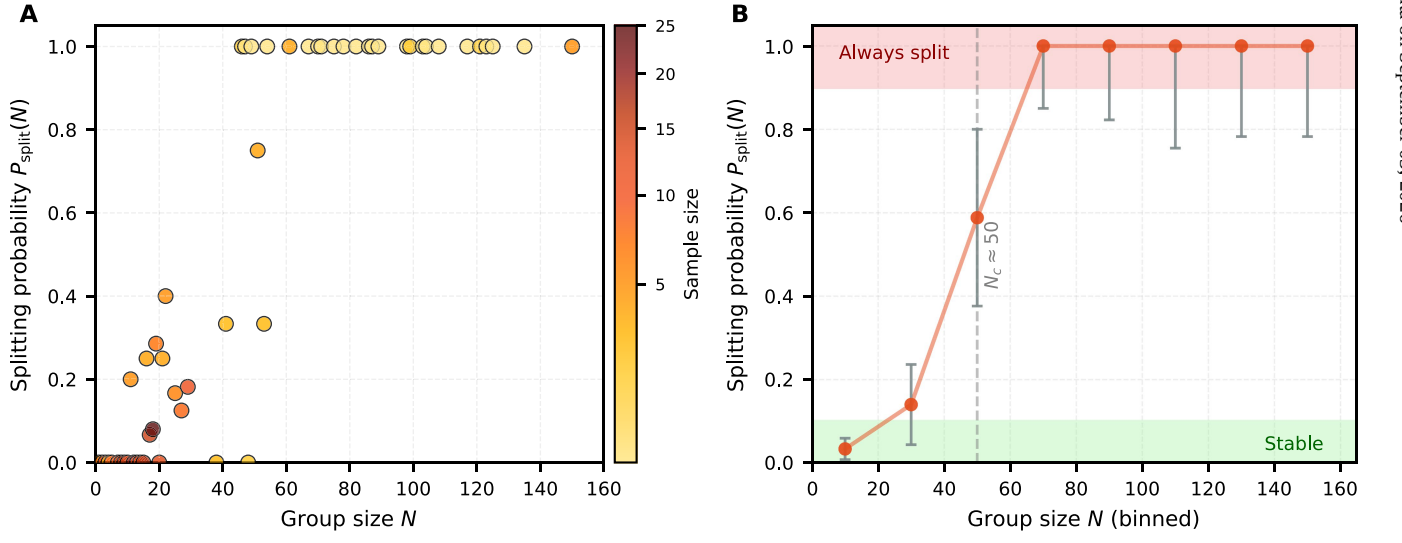


**Fig. 5. Splitting probability in AI agent groups.** (**A**) Splitting probability for individual group sizes observed across all five simulation cases. Each point represents a unique group size, with color indicating the number of observations (sample size). Groups smaller than $N \approx 15$ exhibit near-zero splitting probability (highly stable), while groups larger than $N \approx 60$ always split. (**B**) Binned splitting probability with 20-unit intervals showing the sharp transition from stable to fragmenting regimes. Error bars represent 95 % Wilson score confidence intervals. For $N < 30$, groups are highly stable, while they are highly unstable as soon as $N > 50$. This size-dependent splitting behavior suggests an intrinsic scale limit for stable group formation in AI agent societies.

**Table 1. Specific model versions used in experiments.** Complete list of all LLM model versions analyzed in our experiments.

| Model name | Model version |
|---|---|
| Claude 3.5 Sonnet | claude-3-5-sonnet-20240620 |
| Claude 3 Haiku | claude-3-haiku-20240307 |
| Claude 3 Opus | claude-3-opus-20240229 |
| Claude 3 Sonnet | claude-3-sonnet-20240229 |
| Claude 2.0 | claude-2.0 |
| GPT-3.5 Turbo | gpt-3.5-turbo-1106 |
| GPT-4 | gpt-4-0613 |
| GPT-4o | gpt-4o-2024-05-13 |
| GPT-4 Turbo | gpt-4-turbo-2024-04-09 |
| Llama 3 70B | meta-llama-3-70b-instruct |

The equilibrium value of the magnetization in the CW model, which is reached from any initial configuration, is given by the self-consistency equation

$$m^* = \tanh(\beta m^*) \tag{2}$$

The solution of this equation depends on the value of the inverse temperature β (which corresponds, in the opinion dynamics framework, to the majority force). For $\beta < 1$, the only solution is $m^* = 0$, while for $\beta > 1$, $m^* = 0$ is an unstable solution, while two new solutions $\pm m^*(\beta)$ appear. The value $\beta_c = 1$ is the critical point of a second-order phase transition. This means that as soon as $\beta > 1$, $m^*$ grows gradually from zero with β, tending to $m^* = 1$ for large values of β, while following approximately the equation (see the Supplementary Materials)

$$m^* = \sqrt{3(\beta - 1)}$$

for β close to 1.

### Critical group size

For $\beta > \beta_c$, the equilibrium value of the magnetization is $|m^*| > 0$, but this does not mean that coordination ($|m| = 1$) is reached. The magnetization fluctuates over time around $m^*$, and the amplitude of fluctuations decreases with the number of agents $N$. Full coordination is reached when a fluctuation leads from $m^*$ to $|m| = 1$. Therefore, the ability to coordinate crucially depends on the interplay between the value of $m^*$ and $N$. If $N$ is not too "large" (in a sense specified below), then normal fluctuations are sufficient to rapidly lead to a fully ordered configuration with $|m| = 1$. In this case, the time to coordinate grows logarithmically with $N$. On the other hand, if $N$ is too large, coordination can be achieved only when a large and extremely rare fluctuation around $m^*$ occurs. The average time for this rare fluctuation grows exponentially with $N$, and hence coordination becomes practically unreachable in reasonable times. Since the equilibrium collective opinion $m^*$ is determined by β, this implies that a crossover value $\beta_t(N)$ separates the two regimes.

To get a more precise picture, we need to study the stochastic differential equation governing the out of equilibrium evolution of $m$. If we are interested in the behavior around the equilibrium magnetization $m^* \approx 1$, we can approximate it with an Ornstein-Uhlenbeck process

$$dm = -(m - m^*)dt + \sqrt{2D}dW(t)$$

with

$$D \approx \frac{2}{N}e^{-2\beta}$$

More detailed computations are reported in the Supplementary Materials. The collective opinion will move randomly around $m^*$ with fluctuations of the order of the variance of $m$, that satisfies

$$\mathrm{Var}[m] \approx D \approx \frac{2}{N}e^{-2\beta} \tag{3}$$

This estimate of the typical fluctuations allows to determine the crossover value $\beta_t$ separating a regime ($\beta > \beta_t$) where fluctuations of the size of the variance lead to coordination, from a regime ($\beta < \beta_t$) where a large rare fluctuation is instead necessary. As shown in the Supplementary Materials, it holds

$$\beta_t \approx \frac{1}{2}\log N \tag{4}$$

Equation 4 is reported in Fig. 3. As expected, $\beta_t > \beta_c = 1$ for $N$ larger than a few units. Note that for what concerns the time to coordinate there is little difference between systems with $\beta_c < \beta < \beta_t$ and "disordered" ones with $\beta < \beta_c$. In the former case, fluctuations occur around a finite value of the magnetization, in the latter around $m^* = 0$. However, in both cases, coordination can be achieved only because of an anomalously large rare fluctuation, requiring exponentially large times to develop. Last, we can estimate the critical group size $N_c$ of AI agents as the size for which the majority force reaches $\beta_t$

$$\beta(N_c) = \frac{1}{2}\log N_c$$

## Supplementary Materials

**This PDF file includes:**

Figs. S1 to S6
Supplementary Text
Tables S1 to S3

## REFERENCES


1. J. Henrich, "The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter," in *The Secret of our Success* (Princeton Univ. Press, 2015).
2. R. I. M. Dunbar, *Grooming, Gossip, and the Evolution of Language* (Harvard Univ. Press, 1996).
3. M. Chen, J. Tworek, H. Jun, Q. Yuan, Henrique Ponde de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. M. Grew, D. Amodei, S. M. Candlish, I. Sutskever, W. Zaremba, Evaluating large language models trained on code. arXiv:2107.03374 [cs.LG] (2021).
4. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, V. Natarajan, Large language models encode clinical knowledge. *Nature* **620** , 172–180 (2023).
5. D. M. Katz, M. J. Bommarito, S. Gao, P. Arredondo, Gpt-4 passes the bar exam. *Philos. Trans. R. Soc. A* **382**, 20230254 (2024).
6. S. T. Aroyehun, L. Malik, H. Metzler, N. Haimerl, A. di Natale, D. Garcia, LEIA: Linguistic embeddings for the identification of affect. *EPJ Data Sci.* **12**, 52 (2023).
7. D. A. Boiko, R. MacKnight, B. Kline, G. Gomes, Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).

8. B. Romera-Paredes, M. Barekatain, A. Novikov, M. Balog, M. P. Kumar, E. Dupont, F. J. R. Ruiz, J. S. Ellenberg, P. Wang, O. Fawzi, P. Kohli, A. Fawzi, Mathematical discoveries from program search with large language models. *Nature* **625**, 468–475 (2024).
9. J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt,S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B.Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Ł. Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Ł. Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. M. Grew, S. M. McKinney, C. M. Leavey, P. M. Millan, J. M. Neil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, Filipe de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, M. Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. J. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, Gpt-4 technical report. arXiv:2303.08774 [cs.CL] (2023).
10. Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu, Z. Yin, S. Dou, R. Weng, W. Qin, Y. Zheng, X. Qiu, X. Huang, Q. Zhang, T. Gui, The rise and potential of large language model based agents: A survey. *Sci. China. Inf. Sci.* **68**, 121101 (2025).
11. A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, T. W. Malone, Evidence for a collective intelligence factor in the performance of human groups. *Science* **330**, 686–688 (2010).
12. Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, C. Wang, "Autogen: Enabling next-gen LLM applications via multi-agent conversations," in *First Conference on Language Modeling* (Conference on Language Modeling; 2024).
13. A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, Diego de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mixtral of experts. arXiv:2401.04088 [cs.LG] (2024).
14. X. Guo, K. Huang, J. Liu, W. Fan, N. Vélez, Q. Wu, H. Wang, T. L. Griffiths, M. Wang, Embodied llm agents learn to cooperate in organized teams. arXiv:2403.12482 [cs.AI] (2024).
15. Z. Liu, Y. Zhang, P. Li, Y. Liu, D. Yang, Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. arXiv:2310.02170 [cs.CL] (2023).
16. Y. Shen, K. Song, X. Tan, D. Li, W. Lu, Y. Zhuang, Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Adv. Neural Inf. Proces. Syst.* **36**, 38154–38180 (2024).
17. N. Johnson, G. Zhao, E. Hunsader, H. Qi, N. Johnson, J. Meng, B. Tivnan, Abrupt rise of new machine ecology beyond human response time. *Sci. Rep.* **3**, 2627 (2013).
18. A. S. Vezhnevets, J. P. Agapiou, A. Aharon, R. Ziv, J. Matyas, E. A. Duéñez-Guzmán, W. A. Cunningham, S. Osindero, D. Karmon, J. Z. Leibo, Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia. arXiv:2312.03664 [cs.AI] (2023).
19. I. D. Couzin, J. Krause, N. R. Franks, S. A. Levin, Effective leadership and decision-making in animal groups on the move. *Nature* **433**, 513–516 (2005).
20. I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J. F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson, N. R. Jennings, E. Kamar, I. M. Kloumann, H. Larochelle, D. Lazer, R. McElreath, A. Mislove, D. C. Parkes, A. 'S.' Pentland, M. E. Roberts, A. Shariff, J. B. Tenenbaum, M. Wellman, Machine behaviour. *Nature* **568**, 477–486 (2019).
21. G. V. Aher, R. I. Arriaga, A. T. Kalai, "Using large language models to simulate multiple humans and replicate human subject studies," in *Proceedings of the 40th International Conference on Machine Learning* (PMLR; 2023), vol. 202, pp. 337–371.
22. L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, D. Wingate, Out of one, many: Using language models to simulate human samples. *Pol. Anal.* **31**, 337–351 (2023).
23. V. Dentella, F. Günther, E. Leivada, Systematic testing of three Language Models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2309583120 (2023).
24. M. Binz, E. Schulz, Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2218523120 (2023).
25. M. Pellert, C. M. Lechner, C. Wagner, B. Rammstedt, M. Strohmaier, AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspect. Psychol. Sci.* **19**, 808–826 (2023).
26. J. W. A. Strachan, D. Albergo, G. Borghini, O. Pansardi, E. Scaliti, S. Gupta, K. Saxena, A. Rufo, S. Panzeri, G. Manzi, M. S. A. Graziano, C. Becchio, Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* **8**, 1285–1295 (2024).
27. I. Grossmann, M. Feinberg, D. C. Parker, N. A. Christakis, P. E. Tetlock, W. A. Cunningham, AI and the transformation of social science research. *Science* **380**, 1108–1109 (2023).
28. C. A. Bail, Can Generative AI improve social science? *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2314021121 (2024).
29. Y. Lu, A. Aleta, C. Du, L. Shi, Y. Moreno, Generative agent-based models for complex systems research: A review. arXiv:2408.09175 [physics.soc-ph] (2024).
30. S. Lai, Y. Potter, J. Kim, R. Zhuang, D. Song, J. Evans, "Position: Evolving AI collectives enhance human diversity and enable self-regulation," in *Proceedings of the 41st International Conference on Machine Learning* (PMLR; 2024).
31. M. Papachristou, Y. Yuan, Network formation and dynamics among multi-LLMs. arXiv:2402.10659 [cs.SI] (2024).
32. S. Chang, A. Chaszczewicz, E. Wang, M. Josifovska, E. Pierson, J. Leskovec, LLMs generate structurally realistic social networks but overestimate political homophily. arXiv:2408.16629 [cs.CY] (2024).
33. A. C. Kozlowski, H. Kwon, J. A. Evans, In silico sociology: Forecasting COVID-19 polarization with large language models. arXiv:2407.11190 [cs.CY] (2024).
34. G. De Marzo, L. Pietronero, D. Garcia, Emergence of scale-free networks in social interactions among large language models. arXiv:2312.06619 [physics.soc-ph] (2023).
35. P. Törnberg, D. Valeeva, J. Uitermark, C. Bail, Simulating social media using large language models to evaluate alternative news feed algorithms. arXiv:2310.05984 [cs.SI] (2023).
36. J. Piao, Z. Lu, C. Gao, F. Xu, Q. Hu, F. P. Santos, Y. Li, J. Evans, Emergence of human-like polarization among large language model agents. arXiv:2501.05171 [cs.SI] (2025).
37. J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software And Technology* (Association for Computing Machinery, New York, NY; 2023), pp. 1–22.
38. Y.-S. Chuang, A. Goyal, N. Harlalka, S. Suresh, R. Hawkins, S. Yang, D. Shah, J. Hu, T. T. Rogers, Simulating opinion dynamics with networks of llm-based agents. arXiv:2311.09618 [physics.soc-ph] (2023).
39. E. Cau, V. Pansanella, D. Pedreschi, G. Rossetti, Language-driven opinion dynamics in agent-based simulations with LLMs. arXiv:2502.19098 [cs.SI] (2025).
40. A. F. Ashery, L. M. Aiello, A. Baronchelli, Emergent social conventions and collective bias in LLM populations. *Sci. Adv.* **11**, eadu9368 (2025).
41. C. Barrie, P. Törnberg, Emergent LLM behaviors are observationally equivalent to data leakage. arXiv:2505.23796 [cs.CL] (2025).
42. L. M. Aplin, D. R. Farine, J. Morand-Ferron, A. Cockburn, A. Thornton, B. C. Sheldon, Experimentally induced innovations lead to persistent culture via conformity in wild birds. *Nature* **518**, 538–541 (2015).
43. L. M. Aplin, B. C. Sheldon, R. McElreath, Conformity does not perpetuate suboptimal traditions in a wild population of songbirds. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 7830–7837 (2017).
44. C. Castellano, S. Fortunato, V. Loreto, Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591–646 (2009).
45. M. Kochmański, T. Paszkiewicz, S. Wolski, Curie–Weiss magnet—A simple model of phase transition. *Eur. J. Phys.* **34**, 1555–1573 (2013).
46. M. Tomasello, M. Carpenter, J. Call, T. Behne, H. Moll, Understanding and sharing intentions: The origins of cultural cognition. *Behav. Brain Sci.* **28**, 675–691 (2005).
47. C. Castellano, Social influence and the dynamics of opinions: The approach of statistical physics. *Manag. Decis. Econ.* **33**, 311–321 (2012).
48. R. J. Glauber, Time-dependent statistics of the Ising model. *J. Math. Phys.* **4**, 294–307 (1963).
49. R. I. Dunbar, Neocortex size as a constraint on group size in primates. *J. Hum. Evol.* **22**, 469–493 (1992).
50. D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding. arXiv:2009.03300 [cs.CY] (2020).
51. S. Zheng, Y. Zhang, Y. Zhu, C. Xi, P. Gao, X. Zhou, Kevin Chen-Chuan Chang, Gpt-fathom: Benchmarking large language models to decipher the evolutionary path towards GPT-4 and beyond. arXiv:2309.16583 [cs.CL] (2023).

52. A. R. DeCasien, S. A. Williams, J. P. Higham, Primate brain size is predicted by diet but not sociality. *Nat. Ecol. Evol.* **1**, 0112 (2017).
53. P. Lindenfors, A. Wartel, J. Lind, 'Dunbar's number' deconstructed. *Biol. Lett.* **17**, 20210158 (2021).
54. M. Casari, C. Tagliapietra, Group size in social-ecological systems. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2728–2733 (2018).
55. B. Gonçalves, N. Perra, A. Vespignani, Modeling users' activity on twitter networks: Validation of dunbar's number. *PLOS ONE* **6**, e22656 (2011).
56. G. De Marzo, D. Garcia, Collective behavior of AI agents: The case of moltbook. arXiv:2602.09270 [physics.soc-ph] (2026).
57. F. Fadaei, J. C. Moran, T. Yasseri, Gender dynamics and homophily in a social network of LLM agents. arXiv:2602.02606 [cs.SI] (2026).
58. S. Bikhchandani, D. Hirshleifer, I. Welch, A theory of fads, fashion, custom, and cultural change as informational cascades. *J. Polit. Econ.* **100**, 992–1026 (1992).

**Acknowledgments**
**Funding:** This project was supported by OpenAI with free API credits under its research program. C.C. acknowledges the PRIN project no. 20223W2JKJ "WECARE," CUP B53D23003880006, financed by the Italian Ministry of University and Research (MUR), Piano Nazionale Di Ripresa e Resilienza (PNRR), Missione 4 "Istruzione e Ricerca"–Componente C2 Investimento 1.1, funded by the European Union–NextGenerationEU.
**Author contributions:** Conceptualization: G.D.M., C.C., and D.G. Methodology: G.D.M., C.C., and D.G. Investigation: G.D.M. Visualization: G.D.M. Supervision: D.G. Writing—original draft: G.D.M. Writing—review and editing: G.D.M., C.C., and D.G. **Competing interests:** The authors declare that they have no competing interests. **Data, code, and materials availability:** All data and code needed to evaluate and reproduce the results in the paper are present in the paper and/or the Supplementary Materials. All code and data generated in this study are publicly available at https://github.com/giordano-demarzo/LLMs-Opinion-Dynamics. The code and data are also archived on Zenodo (https://doi.org/10.5281/zenodo.17911594) to ensure long-term accessibility and reproducibility. This study did not generate new materials.

Submitted 15 July 2025
Accepted 7 July 2026
Published 14 August 2026
10.1126/sciadv.aea6091

Science Advances

# AI agents can coordinate via majority-following beyond human scale

Giordano De Marzo, Claudio Castellano, and David Garcia

*Sci. Adv.* **12** (33), eaea6091. DOI: 10.1126/sciadv.aea6091



*Science Advances* (ISSN 2375-2548) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.