Modeling Multi-Agent LLM Debates Through Damped Ordinary Differential Equations*

Leonardo Mascagni[1,2,†], Andrea Agosta[1,2,†], Mauro Mezzenzana[1,2,†], Giacomo Buonanno[1,2,†] and Francesco Bertolotti[2,3,*,†]

[1]School of Industrial Engineering, Cattaneo University – LIUC, Italy
[2]Intelligence, Complexity, and Technology Lab (ICT Lab), University Cattaneo – LIUC, Italy
[3]Università Cattolica di Milano, Department of Philosophy, L. Gemelli 1 - 20123 Milano, Italy

Abstract

We study the reproducibility of multi-agent LLM debates by repeating the same debate configuration multiple times and representing each run as a trajectory in a low-dimensional opinion space. To quantify cross-run variability, we introduce a positional-dispersion measure over turns and model its evolution with a damped second-order ordinary differential equation. This framework captures contraction, rebound, and persistent disagreement across debate rounds, allowing models to be compared not only by outcome quality but also by the stability of their interaction dynamics. Focusing on Gemma and Llama models across balanced and asymmetric debate settings, we find that reproducibility varies substantially with model family and size: Gemma generally follows narrower and more stable trajectories, whereas Llama retains broader variability and shows greater sensitivity to changes in the agent-to-opinion ratio. These results suggest that reproducibility is a useful additional dimension for evaluating multi-agent LLM systems.

Keywords

LLM, Multi-agent systems, dynamical model, complex systems, differential equation, dumping,

1. Introduction

Multi-agent LLM systems are increasingly employed to deliberate, critique, and refine answers through interaction rather than through a single forward pass [1], with debate protocols, round-table exchanges, and related agentic workflows used both as practical strategies for improving reasoning quality and as experimental settings for studying collective model behavior [2, 3]. Yet these systems are still evaluated mainly through their final output [4]. This leaves in the background a question: when the same debate is repeated under the same initial conditions, does the collective process remain on a reproducible path, or does it bifurcate because of small prompt-level perturbations?

The question is relevant because final-answer quality is not sufficient to characterize the behavior of an interacting population of LLMs. A debate may end in the correct place while traversing very different intermediate paths, and a system may appear stable only because disagreement is suppressed too early. In both cases, an important part of the phenomenon is missed if attention is restricted to the final turn. Recent work on LLM populations, conformity, and opinion dynamics has already shown that collective behavior depends on model family, framing, alignment, and on the very structure of the opinion space [5, 6, 7, 8]. What is still lacking is a compact way to quantify how reproducible those interaction trajectories are across repeated runs.

In this paper, we study the dynamics of debates, repeating the same configuration multiple times for each model, encoding every run as a trajectory in a low-dimensional geometric opinion space, and

WOA 2026, the 27th Workshop From Objects to Agents

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Corresponding author.
†These authors contributed equally.
le21.mascagni@stud.liuc.it (L. Mascagni); an17.agosta@stud.liuc.it (A. Agosta); mmezzenzana@liuc.it (M. Mezzenzana); buonanno@liuc.it (G. Buonanno); fbertolotti@liuc.it (F. Bertolotti)
0000-0003-1274-9628 (F. Bertolotti)

measure cross-run variability through a positional-dispersion statistic over turns. Then, we fit the resulting curves with a second-order ordinary differential equation (ODE), that is, a model involving the quantity itself together with its first and second derivatives over time. This gives a compact description of damping, rebound, and persistent disagreement floors. It also allows models to be compared by how they converge and by how consistently they retrace the same path.

The results show that repeated debates are not uniformly stable, but all of them shows some form of elastic behavior. Smaller dense models preserve broader disagreement floors and react more strongly to changes in the agent-to-opinion ratio. The asymmetric two-agent/three-opinion condition is particularly informative, because the presence of an additional available stance does not have a uniform effect: for some models it stabilizes the debate, whereas for others it keeps alternative trajectories alive.

Hence, this work makes two main contributions. First, it introduces a geometric-dynamical framework for studying repeated debate trajectories in both balanced and asymmetric settings. Second, it shows empirically that reproducibility can be used as an additional comparison dimension for multi-agent LLM systems alongside final-answer metrics. This distinction matters in settings where the stability of the interaction process is relevant, not only its endpoint. Any link between these patterns and model construction should be read as a hypothesis rather than as a direct inference.

The paper proceeds as it follows. Firstly, the background of the research is presented. Then, the methodology is depicted, results presented and discussed, and the conclusions drawn.

2. Background

Multi-agent debate is now one of the main paradigms used to make language-model reasoning interactive rather than purely one-shot [9]. The basic intuition is well established: exposing a model to criticism, alternative arguments, and iterative revision can improve factuality and reasoning quality [10, 11]. For this reason, debate protocols have been adopted not only as practical prompting strategies, but also as experimental settings in which collective model behavior can be observed more directly than in single-response generation [12].

Early theoretical analyses already suggested that debate outcomes depend not only on the intrinsic capability of the agents, but also on the correlation between their responses and on the architecture of the interaction itself [13]. More recent empirical studies have reinforced this point, showing that agent personas, turn-taking rules, prompt design, and aggregation procedures all shape the final collective outcome [14, 15]. Related protocols such as round-table consensus and collaborative peer-review systems point in the same direction: interaction is useful when disagreement carries informative variation, but can become redundant when agents merely restate compatible positions [16, 17].

In this context, it is important to address the tendency of LLMs to conform, flatter, or otherwise move toward socially reinforced positions [18]. Psychological-style experiments have shown that contemporary LLMs display measurable conformity effects, including in cases where the majority opinion is not correct [19]. Instruction tuning may mitigate this tendency in some settings, but does not remove it [20]. This matters directly for multi-agent debate, because the apparent success of interaction may sometimes reflect truth-seeking and sometimes mere social convergence.

The same concern appears in work on sycophancy, bias, and adversarial influence [21]. When agents reinforce one another without preserving substantive disagreement, debate can become less reliable and less efficient, despite involving more interaction [21]. Other studies have shown that multi-agent configurations can dampen or intensify political bias depending on how the debate is structured [14], while recent work on adversarial persuasion in multi-agent debate shows that apparently coherent deliberation can drift toward confidently shared but wrong conclusions [22]. So, interaction is not automatically epistemically beneficial: it can produce correction, but it can also produce premature consensus.

Opinion-simulation studies sharpen this point further. Networks of LLM-based agents often exhibit a built-in pull toward consensus and toward answers perceived as safer, more acceptable, or more coherent with previous statements [5, 6]. From the perspective of this paper, this implies that disagreement

should not be treated as an issue to eliminate immediately but it is itself part of the phenomenon to be explained.

Finally, this work is also related to the broader literature on opinion dynamics and consensus formation [23]. Classical models, such as DeGroot-style [24] or Delphi techniques [25], offer a useful baseline: repeated exchange tends to reduce disagreement and often produces approximately exponential contraction toward a limiting state [7]. Yet LLM interactions do not fit cleanly into this simple picture [26]. Empirical studies suggest that the limiting configuration depends not only on initial conditions, but also on topic, framing, model family, and model-specific biases [7].

Recent work has also shown that the geometry of the opinion space matters [6]. Consensus tends to be stronger when agents choose among a fixed set of explicit options, whereas freer opinion spaces allow more heterogeneous trajectories; likewise, in language-rich debates, argumentative structure itself can determine whether competing views are accepted, rejected, or bypassed [6, 8].

At the same time, newer approaches to opinion dynamics have moved toward richer dynamical descriptions, including differential-equation models intended to capture persistence, rebound, and non-trivial collective regimes [27]. Information-theoretic analyses of multi-agent interaction similarly suggest that coordination should be treated as an emergent property of the process itself, not merely as a by-product of isolated responses [15]. Our contribution is positioned at the intersection of these perspectives. We represent repeated debates as trajectories in a low-dimensional geometric opinion space and summarize their cross-run variability through a second-order dynamical description of positional dispersion.

3. Methods

3.1. Debate generation

For each experimental condition, we repeatedly generated debates under controlled initial opinions. The pipeline supports both balanced and asymmetric regimes, where the number of agents may differ from the number of admissible opinions. In the experiments reported here, this included 2–2, 2–3, 3–3, and 3–4 settings. Typical examples were *left hand* vs. *right hand* in the 2–2 case, *Charmander* vs. *Squirtle* with *Bulbasaur* left available in the 2–3 case, and *North*, *South*, and *East* with *West* left free in the 3–4 case. In the balanced regimes, each agent was initialized with one predefined stance. When the agent-to-opinion ratio was different from 1, one admissible option remained initially unassigned and could emerge during the debate as a free alternative. The options were chosen to be as neutral as possible, so that the observed trajectories depended mainly on the interaction dynamics rather than on the semantic attractiveness of a specific stance.

For every run, the orchestration script stored the model name, topic, number of rounds, initial opinion assignment, and full ordered response sequence in a structured log. This preserved the full debate history for the downstream parsing, geometric encoding, and aggregation steps described below. The reported curves were generated over a 40-round horizon, with each configuration typically repeated 50 times.

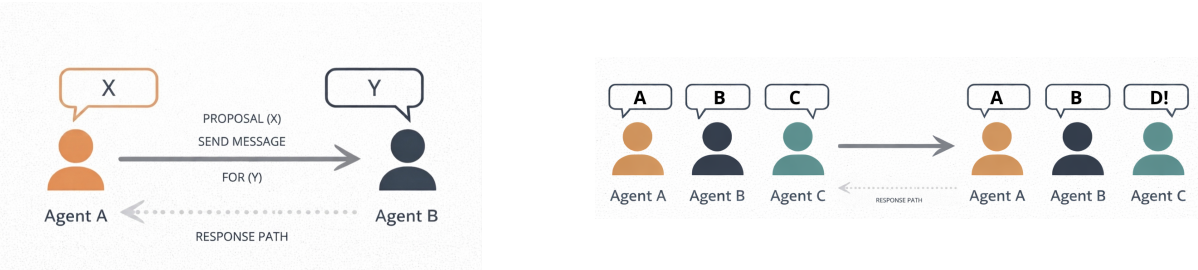

Figure 1: Schematic examples of the debate setup: two-agent interaction (left) and three-agent interaction (right).

3.2. Opinion definition and response parsing

Each debate log was parsed to recover the model, topic, initial opinions, and ordered response sequence. Stance assignment was performed through a second-pass semantic classification rather than simple string matching, so that labels reflected the interpreted meaning of each response.

In debates starting from two explicit opinions, responses were mapped to four classes: A and B for the two assigned stances, C for a coherent third alternative, and D for out-of-context or unusable replies. This covers both the 2–2 and 2–3 regimes, where the third option is respectively latent or explicit but initially unassigned.

In debates starting from three explicit opinions, responses were mapped to five classes: A, B, and C for the assigned stances, D for a coherent fourth alternative, and E for out-of-context or unusable replies. This covers both the 3–3 and 3–4 regimes. The result is a symbolic sequence over $\{A, B, C, D\}$ or $\{A, B, C, D, E\}$ that preserves the distinction between assigned positions, constructive alternatives, and off-topic behavior.

3.3. Geometric representation of debate trajectories

A geometric representation was chosen to match the number of meaningful opinion categories available in each regime. When the debate contains three meaningful options overall, as in the 2-2 and 2-3 settings, responses are embedded in a triangular opinion space. In this case, A and B are the two assigned initial opinions, while C denotes the constructive or latent third option. Let $n_A(t)$, $n_B(t)$, $n_C(t)$, and $n_D(t)$ denote the cumulative number of responses classified as A, B, C, and D up to turn t, respectively. Since D is out of context, it does not contribute to the geometric position. The triangle vertices are defined as

$$A = (0, 0), \qquad B = (1, 0), \qquad C = \left(0.5, \frac{\sqrt{3}}{2}\right),$$

and, with

$$N_2(t) = n_A(t) + n_B(t) + n_C(t),$$

the debate position at turn t is

$$x_2(t) = \frac{n_B(t) + 0.5\, n_C(t)}{N_2(t)}, \qquad y_2(t) = \frac{\sqrt{3}}{2} \frac{n_C(t)}{N_2(t)}.$$

Graphically, this case is shown as a triangle whose lower vertices correspond to the two assigned initial opinions and whose upper vertex corresponds to the third available option. In the 2-2 condition, that upper vertex represents a constructive alternative not fixed in advance; in the 2-3 condition, it represents the explicit but initially unassigned third opinion. Debate trajectories are plotted as successive barycentric positions inside the triangle. Responses labelled D are excluded from the coordinates and treated only as a diagnostic marker of out-of-context behavior.

When the debate contains four meaningful options overall, as in the 3-3 and 3-4 settings, the analysis follows a square representation. The three assigned initial opinions occupy three vertices of the square, while the remaining vertex represents the fourth available option. Let $n_A(t)$, $n_B(t)$, $n_C(t)$, $n_D(t)$, and $n_E(t)$ denote the cumulative number of responses classified as A, B, C, D, and E up to turn t, respectively. The square vertices are

$$A = (0, 0), \qquad B = (1, 0), \qquad C = (1, 1), \qquad D = (0, 1),$$

and, with

$$N_3(t) = n_A(t) + n_B(t) + n_C(t) + n_D(t),$$

the position at turn t is

$$x_3(t) = \frac{n_B(t) + n_C(t)}{N_3(t)}, \qquad y_3(t) = \frac{n_C(t) + n_D(t)}{N_3(t)}.$$

Graphically, this case is shown as a square in which A lies at the bottom-left corner, B at the bottom-right, C at the top-right, and D at the top-left. In the 3-3 condition, D acts as a constructive fourth alternative that is not initially assigned; in the 3-4 condition, D corresponds to the explicit fourth option left free at initialization. The trajectory starts from the center of the square, $(0.5, 0.5)$, and is updated round by round using the cumulative barycenter of A, B, C, and D only. Responses labelled E remain outside the geometric space because they mark out-of-context behavior; in the corresponding visualization, they affect the color coding of the plotted points but not their coordinates. More generally, both mappings convert a debate run into a discrete trajectory: the triangular representation captures regimes with three meaningful options, whereas the square representation captures regimes with four meaningful options.

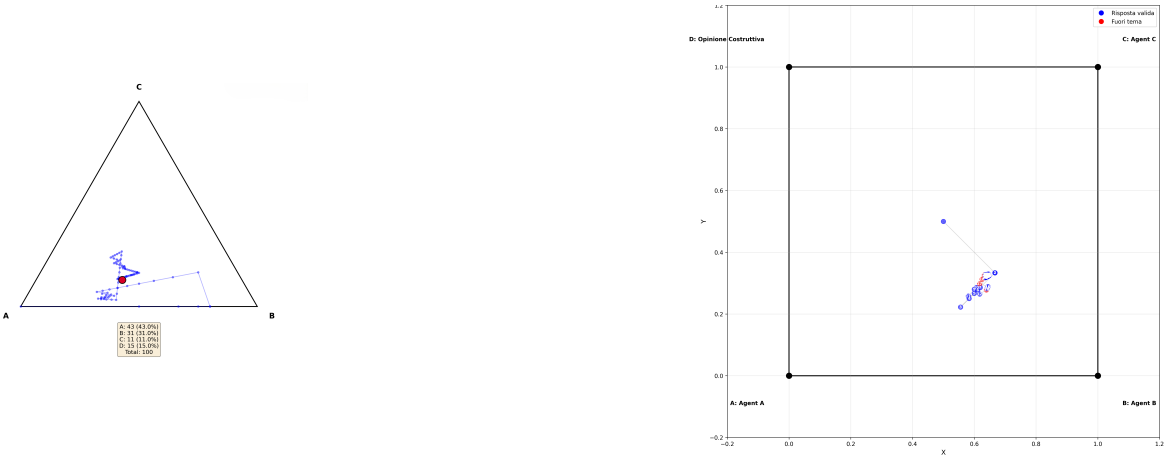

Figure 2: Examples of debate trajectories in the geometric opinion spaces: triangular representation (left) and square representation (right).

3.4. Repeated debates and positional dispersion

To quantify the stability of model behavior, the same debate configuration was repeated multiple times for each model. For every run, the debate trajectory was computed turn by turn in the corresponding geometric opinion space: a triangle when three meaningful options were available and a square when four meaningful options were available. Then, for each turn t, the empirical distribution of trajectory positions across repeated debates was used to compute the standard deviation of the x- and y-coordinates:

$$S[X(t)], \qquad S[Y(t)].$$

From these two quantities, we defined the radial positional dispersion, denoted by $\eta(t)$, as

$$\eta(t) = \sqrt{S[X(t)]^2 + S[Y(t)]^2}.$$

This scalar observable provides a compact measure of how dispersed the debate trajectories are across repeated runs at turn t. Low values indicate that repeated debates generated by the same model follow similar paths, whereas high values indicate greater behavioral variability. The time series

$$s(t) = \eta(t)$$

is the main observable analyzed in this work. In the diagnostic plots, the same quantity is sometimes labeled simply as `sr`, used as a shorthand for η.

3.5. Model-level aggregation

Debates were grouped by model after a normalization step on model names, so that runs belonging to the same model family and size could be aggregated consistently even when logging conventions differed

across files. For each model and each turn, the aggregation procedure produced the mean trajectory position together with the dispersion statistics $S[X(t)]$, $S[Y(t)]$, and $\eta(t)$. When available, model-level benchmark metadata such as MMLU-Pro scores were also associated with the aggregated results for descriptive comparison. Because the present study is centered on repeated-run reproducibility rather than benchmark prediction, these external scores are used only as descriptive anchors and not as optimization targets.

3.6. Second-order dynamical fitting

To obtain a compact dynamical description of the evolution of positional dispersion, we modeled the empirical time series $s(t) = \eta(t)$ through the following second-order ordinary differential equation (ODE):

$$s''(t) + a\,s'(t) + b\,s(t) = c.$$

The aim of this step was to summarize the temporal evolution of dispersion through three interpretable coefficients (a, b, c) and two fitted initial conditions. For each model, the unknown parameter vector was defined as

$$\theta = (a, b, c, s_0, v_0),$$

where

$$s_0 = s(t_0), \qquad v_0 = s'(t_0)$$

represent the fitted initial state at the selected starting turn t_0.

The starting turn t_0 was chosen separately for each model in order to focus the fit on the portion of the trajectory considered most informative for the dynamical regime under analysis. This choice made it possible to exclude very early transient behavior when necessary, but it also means that the fitted coefficients should be read as local descriptive summaries of the observed regime rather than as globally identified physical constants.

3.7. Numerical solution and parameter estimation

The second-order equation was rewritten as the first-order system

$$\begin{cases} s'(t) = v(t), \\ v'(t) = c - a\,v(t) - b\,s(t). \end{cases}$$

For each candidate parameter vector θ, the system was numerically integrated over the observed turn interval using the empirical turn indices as evaluation points. Let $\{t_0, t_0 + 1, \dots, T\}$ be the turns included in the fit, and let $\hat{s}(t; \theta)$ denote the numerical solution of the ODE. Residuals were defined as the difference between the fitted trajectory and the empirical dispersion values:

$$r_t(\theta) = \hat{s}(t; \theta) - s_{\text{obs}}(t), \qquad t = t_0, \dots, T.$$

Parameter estimation was performed by minimizing the sum of squared residuals through bounded nonlinear least squares:

$$\min_{\theta} \sum_{t=t_0}^{T} r_t(\theta)^2.$$

To reduce sensitivity to initialization, the optimization was repeated from 25 random initial guesses. The final parameter set retained for each model was the one achieving the lowest residual cost among all optimization runs. The search bounds allowed both positive and negative values for the dynamical coefficients and for the initial velocity, while constraining the optimization to a numerically stable range.

For each fitted model, we recorded the estimated coefficients (a, b, c), the selected starting turn t_0, the fitted initial conditions (s_0, v_0), and the mean squared error (MSE) over the fitted points. In addition,

diagnostic plots reported the coefficient of determination R^2 between the empirical η sequence and the fitted trajectory evaluated at the observed turns. This gives a standard goodness-of-fit summary alongside the residual scale captured by the MSE.

The full pipeline therefore consisted of four consecutive stages: debate generation, semantic classification, geometric-statistical aggregation, and second-order dynamical fitting. This modular organization preserved the full chain from raw debate text to the final dynamical coefficients used for model comparison.

3.8. Experimental Setup

We evaluated four instruction-following LLM configurations spanning compact dense families. The tested set included Llama 3.2 in 1B and 3B versions and Gemma 3 in 4B and 12B versions. Each model was tested under repeated debates in multiple agent-to-opinion regimes, including the balanced 2-2 and 3-3 settings and the asymmetric 2-3 setting reported in this paper. The debate pipeline fixed the initial stance allocation for each regime, stored the full ordered transcript of each run, and then aggregated the resulting trajectories by normalized model identifier. In the released summaries used here, each debate trajectory is tracked for 40 rounds; the project instructions indicate that configurations were typically repeated 50 times per model-condition pair.

The central experimental choice was to repeat the same debate configuration many times for the same model rather than evaluate a single run per topic. This design isolates reproducibility: if a model repeatedly follows nearly identical geometric trajectories, its dispersion remains low; if small prompt-level fluctuations or revision instabilities accumulate across runs, the dispersion curve broadens. Model-specific benchmark metadata such as MMLU-Pro were associated with the aggregated results when available, not as optimization targets but as descriptive anchors for comparing capability and reproducibility.

4. Results

4.1. Balanced Debate Settings

We begin from the balanced regimes in which each agent is initially associated with exactly one opinion. This includes the two-agent/two-opinion setting (2-2) and the three-agent/three-opinion setting (3-3). These two conditions are especially useful as a baseline because they remove the asymmetry introduced by latent or unassigned alternatives and therefore isolate the intrinsic tendency of the debate dynamics to either preserve disagreement or contract toward a more reproducible trajectory.

In the 2-2 condition, model behavior remained heterogeneous across families. Measuring the initial dispersion at turn 2, the repeated-debate trajectories typically started from a relatively dispersed configuration and then contracted over time, but the rate and magnitude of this contraction varied substantially by model. For instance, `llama3.2:3b` showed the strongest decrease in positional dispersion, with η dropping from about 0.418 at turn 2 to about 0.158 at the final snapshot, while `llama3.2:1b` remained much more variable throughout the interaction and still ended near 0.249. The two Gemma variants occupied an intermediate regime, ending around 0.122 and 0.135, respectively. Table 1 records the final dispersion explicitly together with the fit quality. Overall, the 2-2 setup therefore revealed a clear separation between models that progressively stabilized toward reproducible debate paths and models that preserved a broader spread across repeated runs.

`gemma3:12b` This dense model shows a smooth monotone contraction from about 0.202 to 0.135, with moderate damping ($a \approx 1.05$) and a nonzero disagreement floor ($c/b \approx 0.135$). The curve suggests stable convergence without rebound: the debate contracts, but not toward zero variance. Interpreted against model characteristics, this is compatible with a medium-scale dense open model whose instruction tuning regularizes responses but does not collapse them into an almost deterministic output distribution. In that reading, a captures a moderate resistance to changing the current revision trend, b is strong

enough to pull debates back from divergence, and c remains large enough to preserve a residual spread. Relative to its MMLU-Pro of 60.6, `gemma3:12b` behaves as a reasonably reproducible dense model that still preserves some run-to-run flexibility.

`gemma3:4b` The 4B dense Gemma follows the same qualitative pattern as the 12B model, but with slightly lower final dispersion (≈ 0.122) and a visibly tighter trajectory. The fit remains overdamped and monotone, so in this binary scenario the smaller Gemma does not oscillate like a spring: it simply contracts in a controlled way toward a positive floor. Compared with the 12B variant, the main difference is not a qualitatively different mechanism but a different balance between scale and residual exploratory freedom: the model is still dense and instruction-tuned, yet its smaller parameter count does not translate into larger rebounds. This makes it more stable than the Llama variants despite its lower MMLU-Pro of 43.6.

`llama3.2:1b` This is the least stable model in the balanced binary setting. The trajectory decays only slowly from about 0.352 to about 0.249, and the fit implies a high disagreement floor. Although the damping term is large, the graph does not show strong consensus; it shows monotone relaxation toward a still-broad spread. For this compact dense Llama variant, the most plausible interpretation is that small scale leaves the model more exposed to prompt-level fluctuations and less able to collapse probability mass onto a single revision path. In the language of the ODE, the system can still dissipate motion, but c remains too large relative to b for the debate to settle into a narrow manifold.

`llama3.2:3b` The 3B Llama starts from the highest dispersion of the whole regime (≈ 0.418 at turn 2) but then contracts substantially to about 0.158. This means it is highly sensitive at the beginning of the debate but still capable of settling once interaction unfolds. Compared with the 1B model, the larger compact variant is much more recoverable: it does not begin stable, but it does learn a common direction over rounds. This is one of the clearest places where parameter count matters inside the same dense family: moving from 1B to 3B does not eliminate early exploratory spread, but it improves the model's ability to re-concentrate its output probabilities after interaction. The corresponding 2-2 trajectories discussed above are shown in Fig. 3 and Fig. 4.

Table 1
Summary of repeated-debate dispersion and ODE fit quality for the 2-2 regime.

Model	Turn 2	Final dispersion	Min	Max	MSE	R^2
`gemma3:12b`	0.202	0.135	0.098	0.269	1.1×10^{-5}	0.984
`gemma3:4b`	0.211	0.122	0.122	0.253	4.2×10^{-6}	0.994
`llama3.2:1b`	0.352	0.249	0.249	0.363	1.8×10^{-5}	0.955
`llama3.2:3b`	0.418	0.158	0.158	0.428	1.1×10^{-5}	0.995

4.2. The 3-3 Balanced Setting

The 3-3 condition was more constrained at the beginning of the interaction. Since each of the three agents starts from a distinct assigned opinion, the first aggregated position is exactly centered in the square opinion space for every run, so the turn-1 dispersion is zero by construction. If dispersion is measured from turn 2 onward, however, variability emerges as soon as models begin revising or defending their positions differently across repetitions. Even in this regime, the models split into distinct stability profiles. `llama3.2:1b` reached the broadest final spread, about 0.240, whereas `gemma3:12b`, `gemma3:4b`, and `llama3.2:3b` all remained below 0.09 at the end of the debate. Table 2 records the final dispersion explicitly together with the fit quality. Compared with the 2-2 case, the 3-3 configuration therefore appears more reproducible overall, while still preserving meaningful model-specific differences in how quickly repeated debates diverge from the common symmetric start.

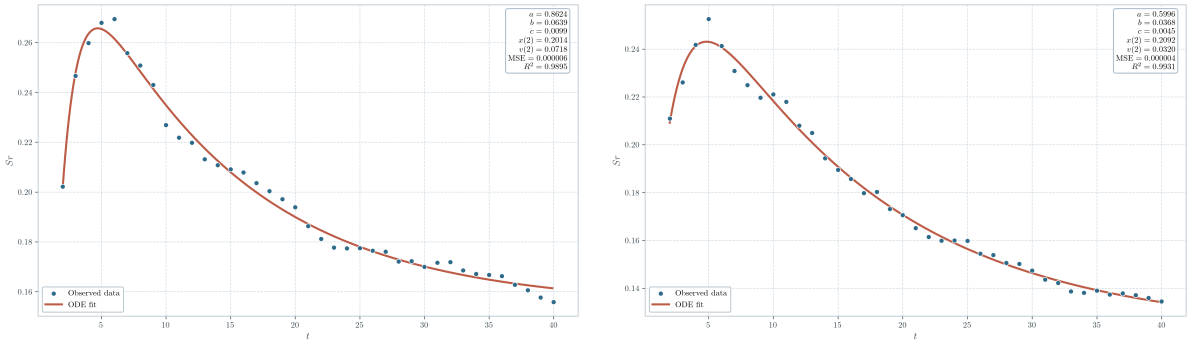

Figure 3: Repeated-debate trajectories and fitted ODE curves in the 2-2 regime for `gemma3:12b` (left) and `gemma3:4b` (right).

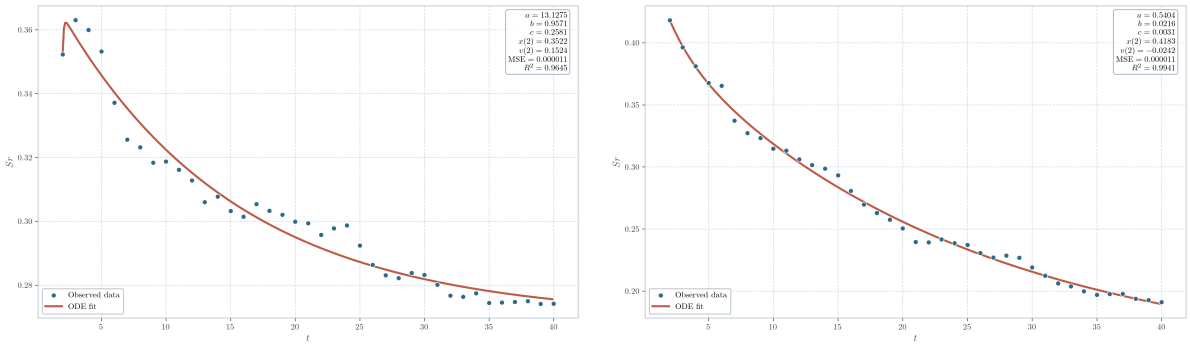

Figure 4: Repeated-debate trajectories and fitted ODE curves in the 2-2 regime for `llama3.2:1b` (left) and `llama3.2:3b` (right).

`gemma3:12b` Here the graph shows a small spring-like correction: dispersion rises slightly, then relaxes to a low final level around 0.049. The negative discriminant suggests a mildly underdamped regime, but the oscillation amplitude stays limited. In other words, `gemma3:12b` does not remain rigid; it corrects itself with small rebounds while staying reproducible. Relative to the broader and less stable `llama3.2:1b` case, this is what one expects from a dense open model whose tuning leaves more probability mass on nearby alternatives: enough elastic return to recover, but not enough to suppress all overshoot.

`gemma3:4b` This is the clearest spring-like trajectory in the whole paper. The curve first drops, then rebounds, peaks again in the middle-late rounds, and only afterward relaxes. The fit is strongly consistent with an underdamped interpretation, so this model is the strongest visual argument for preferring a second-order description over a purely first-order one. Although the final dispersion remains moderate (≈ 0.086), the route taken to get there is visibly oscillatory.

`llama3.2:1b` This remains the outlier of the regime. Even from a perfectly symmetric center, the model quickly opens a large spread and finishes at about 0.240, by far the broadest final dispersion in 3-3. The graph shows that adding one more agent does not regularize the smallest Llama; if anything, the extra interaction channels amplify its instability. In model terms, this is consistent with a small dense model whose output probabilities are not sharply concentrated after alignment, so every new interaction channel creates another opportunity for divergence rather than for coordinated correction.

`llama3.2:3b` In contrast with its 2-2 behavior, the 3B Llama becomes much more disciplined here. The trajectory contracts cleanly toward a low final dispersion of about 0.031, and the fit is sharply overdamped. This suggests that the fully balanced three-opinion geometry stabilizes the model rather

than destabilizing it. Within the dense Llama family, the 3B scale therefore seems sufficient to turn extra interaction structure into a restoring signal instead of a source of persistent spread. The corresponding 3-3 trajectories discussed above are shown in Fig. 5 and Fig. 6.

Table 2
Summary of repeated-debate dispersion and ODE fit quality for the 3-3 regime.

Model	Turn 2	Final dispersion	Min	Max	MSE	R^2
`gemma3:12b`	0.082	0.049	0.047	0.101	5.1×10^{-6}	0.984
`gemma3:4b`	0.086	0.086	0.064	0.098	1.1×10^{-6}	0.992
`llama3.2:1b`	0.142	0.240	0.142	0.270	3.2×10^{-5}	0.935
`llama3.2:3b`	0.081	0.031	0.031	0.091	1.1×10^{-6}	0.996

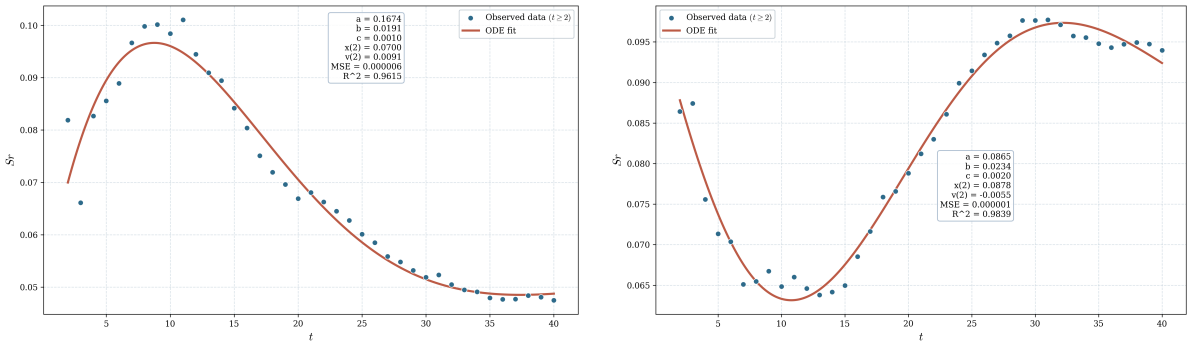

Figure 5: Repeated-debate trajectories and fitted ODE curves in the 3-3 regime for `gemma3:12b` (left) and `gemma3:4b` (right).

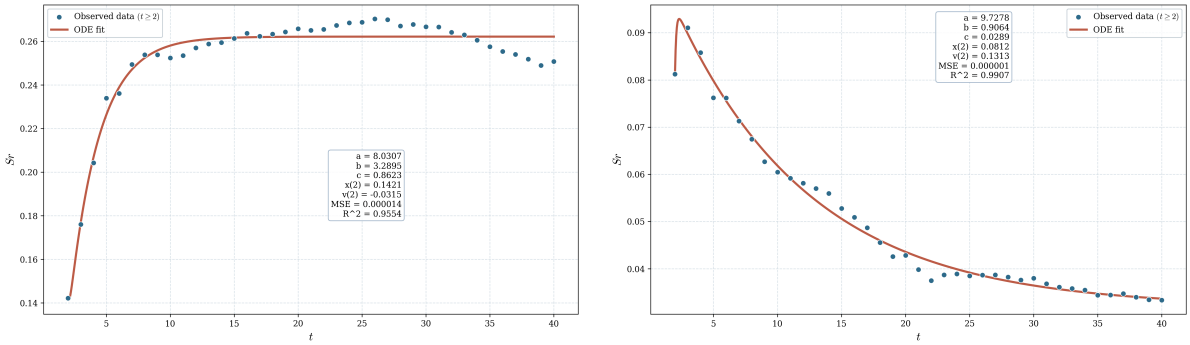

Figure 6: Repeated-debate trajectories and fitted ODE curves in the 3-3 regime for `llama3.2:1b` (left) and `llama3.2:3b` (right).

4.3. The 2-3 Asymmetric Setting

We next consider the asymmetric two-agent/three-opinion setting (2-3), where two opinions are initially assigned to the debaters while a third coherent alternative remains explicitly available but unassigned. Relative to the balanced 2-2 baseline, this regime introduces an additional degree of freedom: repeated debates can now stabilize not only by converging toward one of the two defended positions, but also by drifting toward the latent third option. This makes the regime a useful test of whether models preserve reproducibility even when the interaction space contains a constructive escape from the original binary opposition.

The 2-3 condition still produced substantial differences across models, but the ranking of stability was not identical to the balanced case. At one extreme, `llama3.2:3b` contracted strongly, from about

0.248 at turn 2 to about 0.055 at the end, whereas `llama3.2:1b` and `gemma3:12b` retained a much broader spread, with final dispersions around 0.151 and 0.131, respectively. `gemma3:4b` remained in an intermediate regime, ending close to 0.100. Table 3 summarizes the main dispersion statistics and fit quality. Overall, the 2-3 regime confirms that adding an explicit free alternative does not affect all dense models in the same way, but instead amplifies model-specific differences in how the debate explores a larger opinion space before settling.

`gemma3:12b` The 12B dense Gemma is more sensitive to the free third option than the other dense models considered here. The fit is strongly overdamped, but the final floor remains relatively high (≈ 0.131). In this regime, the extra option remains active across repeated runs instead of disappearing quickly. A possible interpretation is that this family keeps more mass on plausible alternatives, which would be consistent with the larger c/b, but the coefficient alone does not establish that explanation.

`gemma3:4b` This model contracts substantially relative to its starting point, from about 0.271 to about 0.100, but the restoring term is very small and the inferred equilibrium is therefore harder to interpret. The graph is more informative than the coefficients here: it shows slow relaxation toward a nonzero plateau rather than collapse to zero. The third option is partly absorbed, but it continues to contribute to cross-run spread. This is consistent with weaker concentration than in the larger or more heavily post-trained families, though the data do not isolate the source of that difference.

`llama3.2:1b` This model improves relative to its 2-2 and 3-3 behavior, but it remains visibly broad, finishing near 0.151. So the third opinion does help reduce polarization, but not enough to move the model into the more stable range reached here by `llama3.2:3b`. The smallest compact Llama can exploit the extra option, but only partially. Again, the likely bottleneck is not the absence of damping but the persistence of a broad effective output distribution, which keeps the forcing term high relative to the available elastic correction.

`llama3.2:3b` This is the model that benefits the most from the asymmetric setup. Compared with 2-2, its variance falls much faster and to a much lower final level (≈ 0.055). For the 3B Llama, the larger opinion space is associated with lower spread rather than higher spread. Within this family, that contrast with the 1B model is consistent with a scale effect, though the mechanism remains hypothetical.

The corresponding 2-3 trajectories discussed above are shown in Fig. 7 and Fig. 8.

Table 3
Summary of repeated-debate dispersion and ODE fit quality for the 2-3 regime.

Model	Turn 2	Final dispersion	Min	Max	MSE	R^2
`gemma3:12b`	0.222	0.131	0.119	0.232	1.2×10^{-5}	0.964
`gemma3:4b`	0.271	0.100	0.100	0.279	9.1×10^{-6}	0.993
`llama3.2:1b`	0.305	0.151	0.151	0.342	3.6×10^{-5}	0.977
`llama3.2:3b`	0.248	0.055	0.055	0.250	4.0×10^{-6}	0.995

The second-order fit provided a compact summary of these temporal patterns across both balanced and asymmetric regimes. In the 2-2 condition, the fitted trajectories achieved low mean squared errors for all models, ranging from approximately 6.3×10^{-7} to 1.8×10^{-5}, while R^2 ranged from about 0.66 to 0.99. In the 3-3 condition, the fit quality remained similarly high, with MSE values between about 2.1×10^{-7} and 3.2×10^{-5} and R^2 values between about 0.88 and 1.00. In the asymmetric 2-3 condition, the fit was again consistently accurate, with MSE between about 4.6×10^{-8} and 3.6×10^{-5} and R^2 between about 0.90 and 0.99. These values indicate that the dispersion curves are structured enough to be summarized with a small number of dynamical parameters, including when the agent-to-opinion ratio differs from one.

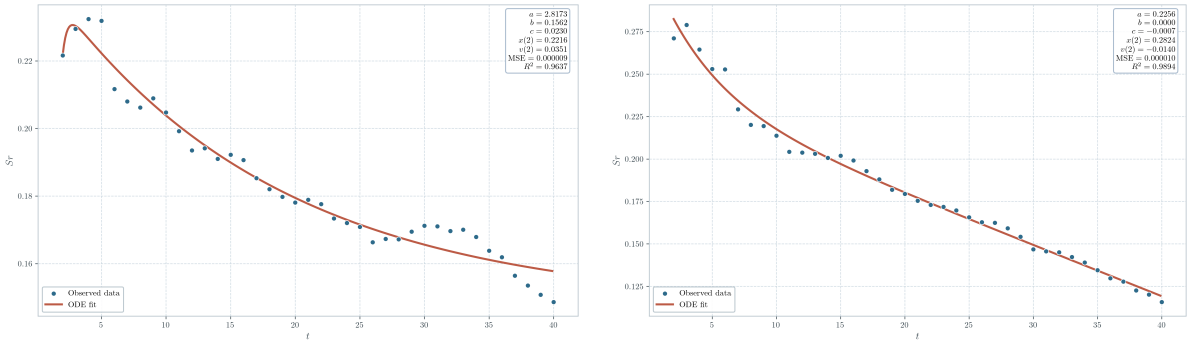

Figure 7: Repeated-debate trajectories and fitted ODE curves in the 2-3 regime for `gemma3:12b` (left) and `gemma3:4b` (right).

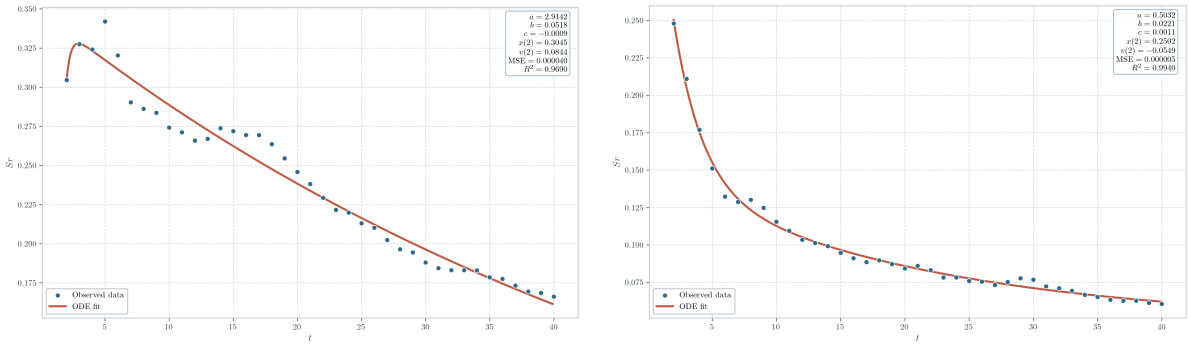

Figure 8: Repeated-debate trajectories and fitted ODE curves in the 2-3 regime for `llama3.2:1b` (left) and `llama3.2:3b` (right).

4.4. A Free Fourth Option

We finally consider the 3-4 regime, where three debaters start from three assigned opinions while a fourth coherent option remains available but initially unassigned. In the current version of the dataset used for this draft, the refitted outputs available for this condition concern the two Gemma models only, so the discussion here is intentionally limited to that family. Even within this narrower comparison, the contrast is informative: `gemma3:12b` remains relatively compact and gradually re-concentrates, whereas `gemma3:4b` broadens markedly and settles on a much higher disagreement floor. Table 4 summarizes the corresponding dispersion statistics and fit quality.

`gemma3:12b` In the 3-4 regime, the 12B Gemma starts from a relatively low dispersion at turn 2 ($\eta \approx 0.045$), rises only modestly to a peak near 0.064, and then gradually contracts to a final value around 0.031. Empirically, this is the more reproducible of the two available Gemma runs in this setting. The fitted coefficient b is close to zero and slightly negative, so the ODE should be read here mainly as a compact descriptive fit rather than as evidence of a strong restoring regime. What matters most is the observed curve itself: the free fourth option does not trigger sustained divergence, and the model slowly re-concentrates over time.

`gemma3:4b` The 4B Gemma behaves quite differently. Starting from $\eta \approx 0.065$ at turn 2, it expands almost continuously through the early and middle part of the debate, reaches a maximum close to 0.140, and still ends at a high final spread of about 0.127. Unlike the 12B variant, this model does not substantially reabsorb the extra freedom introduced by the fourth available option. The trajectory is therefore consistent with a high disagreement floor: the debate remains exploratory, and the added option continues to support noticeable run-to-run variability even late in the interaction. The corresponding 3-4 trajectories discussed above are shown in Fig. 9.

Even in this restricted 3-4 comparison, the fit quality remains good, with MSE between about 1.6×10^{-6} and 4.4×10^{-6} and R^2 between about 0.96 and 0.97. This suggests that the same low-dimensional dynamical description remains informative even when the opinion space is enlarged and only a subset of models is available for analysis.

Table 4
Summary of repeated-debate dispersion and ODE fit quality for the 3-4 regime, restricted to the available Gemma models.

Model	Turn 2	Final dispersion	Min	Max	MSE	R^2
`gemma3:12b`	0.045	0.031	0.031	0.064	1.6×10^{-6}	0.959
`gemma3:4b`	0.065	0.127	0.065	0.140	4.4×10^{-6}	0.974

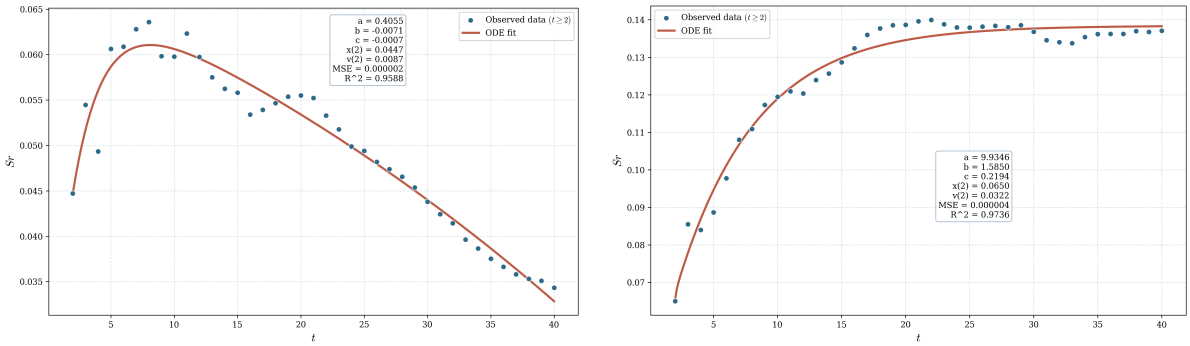

Figure 9: Repeated-debate trajectories and fitted ODE curves in the 3-4 regime for `gemma3:12b` (left) and `gemma3:4b` (right).

4.5. Interpretation of the dynamical coefficients

Interpreting the fitted ODE as a damped spring is a useful descriptive shortcut. In the equation

$$s''(t) + a\,s'(t) + b\,s(t) = c,$$

the coefficient a plays the role of damping, b acts as a restoring term, and c sets a nonzero forcing level. In behavioral terms, high a means that changes in dispersion are quickly dissipated. The coefficient b represents return toward a common path, while c represents persistent reinjection of spread. When $b > 0$, the implied disagreement floor is approximately $s^\star = c/b$; when $a^2 - 4b < 0$, the system is underdamped and may show rebounds or overshoots, whereas $a^2 - 4b > 0$ corresponds to an overdamped or monotone relaxation. No single coefficient should be read as a general quality marker: some models have high damping and still settle at a substantial disagreement floor, while others stay close to zero because both the forcing term and the observed variance are small.

Across the tested families, final reproducibility appears to track capability more clearly than any single ODE coefficient. In our sample, the Pearson correlation between MMLU-Pro and final dispersion is about -0.96 in the 2-2 regime, -0.81 in 3-3, and -0.78 in 2-3. By contrast, the coefficient a alone does not show a stable monotonic relation with MMLU-Pro. This suggests that stronger models are not simply "more inertial" or "more damped"; rather, they tend to combine lower forcing, lower disagreement floors, and more consistent revision behavior.

A useful working hypothesis is that the three coefficients may reflect different aspects of model construction. Strong post-training, with RL-style alignment, is one possible explanation for larger effective a, because it could make later responses less sensitive to small prompt-level perturbations. Architectures or training recipes that keep several plausible answer modes alive might instead appear as smaller b or larger c, which would be consistent with a weaker pull back toward a single path or a

higher disagreement floor. Within our sample, the dense models considered here often retain more spread and exhibit higher disagreement floors. Parameter count also seems to matter within families: moving from Llama 1B to 3B improves re-concentration after early spread. These are pattern-level interpretations, not controlled causal attributions.

Within the Llama family, the observed standard deviation tends to decay toward a stable plateau, especially for the 3B variant. A possible interpretation is that the larger model re-concentrates more effectively after early spread. The contrast between Llama 1B and 3B is consistent with that reading, but it does not by itself identify the mechanism.

We therefore treat the mapping from (a, b, c) to model traits as an interpretive bridge rather than a causal identification. Some families expose their architecture and training choices publicly, whereas others do not, and multiple ingredients may push the coefficients in the same direction. The safest claim is descriptive: lower floors and narrower trajectories are associated with some families more than others, while the explanation remains open.

5. Conclusion

We introduced a geometric-dynamical framework for analyzing repeated multi-agent debates under controlled initial opinions. By mapping each run to a trajectory in an opinion space and tracking cross-run dispersion over time, we showed that reproducibility can be treated as a measurable property of multi-agent LLM interaction rather than as a by-product of final-answer accuracy.

Across the tested regimes, repeated debates often contracted toward stable trajectories, but the extent of that contraction depended strongly on the model family. Within our sample, the smaller dense models retained broader disagreement floors and were more sensitive to changes in the interaction setup, highlighting systematic differences in how models explore and stabilize within the opinion space. The 2-3 regime was informative because the additional option did not affect all models in the same way.

The second-order ODE fit provides a compact descriptive summary of these patterns through damping, elastic return, and persistent forcing. Even in nearly flat trajectories, the combination of geometric dispersion and low-dimensional modeling helps distinguish systems that retrace the same debate path reliably from those that do not. The framework is meant to complement standard evaluation, not replace it.

These results suggest that multi-agent debate should be evaluated not only through final answers, but also through the stability of the trajectory that leads to them. Repeated-run statistics can distinguish systems that converge robustly from systems that remain sensitive to small perturbations or preserve a broader space of alternatives. That distinction may matter in collaborative, deliberative, or decision-support settings where reproducibility is important.

In this paper, the framework is descriptive rather than causal. It can help guide model choice by separating systems that quickly contract toward a narrow path from systems that preserve more variation across runs. It can also be used to formulate hypotheses about model construction, but the fitted coefficients alone do not justify stronger claims about training choices or architecture.

References

[1] A. Bilal, M. A. Mohsin, M. Umer, M. A. K. Bangash, M. A. Jamshed, Meta-thinking in llms via multi-agent reinforcement learning: A survey, arXiv preprint arXiv:2504.14520 (2025).

[2] J. W. Burton, E. Lopez-Lopez, S. Hechtlinger, Z. Rahwan, S. Aeschbach, M. A. Bakker, J. A. Becker, A. Berditchevskaia, J. Berger, L. Brinkmann, et al., How large language models can reshape collective intelligence, Nature human behaviour 8 (2024) 1643–1655.

[3] R. Willis, J. Zhao, Y. Du, J. Z. Leibo, Evaluating collective behaviour of hundreds of llm agents, arXiv preprint arXiv:2602.16662 (2026).

[4] J. De Curtò, I. De Zarzà, Llm-driven social influence for cooperative behavior in multi-agent systems, IEEE Access (2025).

[5] 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, 2024. ArXiv preprint.
[6] P. Cisneros-Velarde, On the principles behind opinion dynamics in multi-agent systems of large language models, 2024. ArXiv preprint.
[7] I. Yazici, M. Kayaalp, S. Taga, A. H. Sayed, Opinion consensus formation among networked large language models, 2026. Preprint.
[8] E. Cau, V. Pansanella, D. Pedreschi, G. Rossetti, Language-driven opinion dynamics in agent-based simulations with llms, 2025. ArXiv preprint.
[9] H. Wu, Z. Li, L. Li, Can llm agents really debate? a controlled study of multi-agent debate in logical reasoning, arXiv preprint arXiv:2511.07784 (2025).
[10] G. Irving, P. Christiano, D. Amodei, Ai safety via debate, arXiv preprint arXiv:1805.00899 (2018).
[11] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, Improving factuality and reasoning in language models through multiagent debate, in: Proceedings of the 41st International Conference on Machine Learning, volume 235 of *Proceedings of Machine Learning Research*, 2024, pp. 11733–11763.
[12] Y. Zhang, X. Yang, S. Feng, D. Wang, Y. Zhang, K. Song, Can llms beat humans in debating? a dynamic multi-agent framework for competitive debate, arXiv preprint arXiv:2408.04472 (2024).
[13] A. Estornell, Y. Liu, Multi-llm debate: Framework, principals, and interventions, in: Advances in Neural Information Processing Systems 37 (NeurIPS 2024), 2024.
[14] A. Bandaru, F. Bindley, T. Bluth, N. Chavda, B. Chen, E. Law, Revealing political bias in llms through structured multi-agent debate, 2025. ArXiv preprint.
[15] C. Riedl, Emergent coordination in multi-agent language models, 2026. Published as a conference paper at ICLR 2026.
[16] J. Chen, S. Saha, M. Bansal, Reconcile: Round-table conference improves reasoning via consensus among diverse llms, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 7066–7085. URL: https://aclanthology.org/2024.acl-long.381/. doi:10.18653/v1/2024.acl-long.381.
[17] Z. Xu, S. Shi, B. Hu, J. Yu, D. Li, M. Zhang, Y. Wu, Towards reasoning in large language models via multi-agent peer review collaboration, arXiv preprint arXiv:2311.08152 (2023).
[18] A. Taubenfeld, Y. Dover, R. Reichart, A. Goldstein, Systematic biases in llm simulations of debates, in: Proceedings of the 2024 conference on empirical methods in natural language processing, 2024, pp. 251–267.
[19] X. Zhu, C. Zhang, T. Stafford, N. Collier, A. Vlachos, Conformity in large language models, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, 2025.
[20] Z. Weng, G. Chen, W. Wang, Do as we do, not as you think: the conformity of large language models, in: The Thirteenth International Conference on Learning Representations, 2025.
[21] P. Pitre, N. Ramakrishnan, X. Wang, CONSENS AGENT: Towards efficient and effective consensus in multi-agent LLM interactions through sycophancy mitigation, in: Findings of the Association for Computational Linguistics: ACL 2025, 2025.
[22] I. Kraidia, et al., When collaboration fails: Persuasion-driven adversarial influence in multi-agent large language model debate, Scientific Reports (2026).
[23] J. A. Hołyst, K. Kacperski, F. Schweitzer, Social impact models of opinion dynamics, Annual Reviews Of Computational PhysicsIX (2001) 253–273.
[24] M. H. DeGroot, Reaching a consensus, Journal of the American Statistical association 69 (1974) 118–121.
[25] N. Dalkey, O. Helmer, An experimental application of the delphi method to the use of experts, Management science 9 (1963) 458–467.
[26] F. Bertolotti, L. Mari, An llm-based delphi study to predict genai evolution, arXiv preprint arXiv:2502.21092 (2025).
[27] C. Gong, Y. Jiang, H. Li, R. Sun, J. Zhang, T. Gu, L. Pan, L. Lü, Advancing opinion dynamics modeling with neural diffusion-convection-reaction equation, 2026. Preprint.