

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. Across Gemma, Llama, gpt-oss, and Gemini configurations, we find that reproducibility varies substantially across model families and scales: Gemini and the two gpt-oss models remain in a very low-dispersion regime, whereas smaller dense models retain broader variability and show 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, while not by itself implying higher accuracy or better reasoning.

Keywords

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

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, 2], 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 [3, 4]. Yet these systems are still evaluated mainly through their final output [5]. 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 [6, 7, 8, 9]. 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 by repeating the same configuration multiple times for each model, encoding every run as a trajectory in a low-dimensional geometric opinion space, and measuring cross-run variability through a positional-dispersion statistic over turns. We then fit the

WOA 2026, the 27th Workshop From Objects to Agents

*Corresponding author.

†These authors contributed equally.

EMAIL: 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)

ORCID: 0000-0003-1274-9628 (F. Bertolotti)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

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. We use it as a descriptive summary rather than as evidence that the underlying process is literally governed by this equation [10].

The results show that repeated debates are not uniformly stable, and that several dispersion curves exhibit contraction, rebound, or persistent disagreement floors. Gemini and the gpt-oss models remain close to narrow, highly reproducible trajectories across regimes, whereas 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 follows. We first present the background, then describe the methodology, report and discuss the results, and draw conclusions.

2. Background

Multi-agent debate is now one of the main paradigms used to make language-model reasoning interactive rather than purely one-shot [11]. The basic intuition is well established: exposing a model to criticism, alternative arguments, and iterative revision can improve factuality and reasoning quality [12, 13]. 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 [14].

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 [15]. 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 [16, 17]. 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 [18, 19].

In this context, it is important to address the tendency of LLMs to conform, flatter, or otherwise move toward socially reinforced positions [20]. Psychological-style experiments have shown that contemporary LLMs display measurable conformity effects, including in cases where the majority opinion is not correct [21]. Instruction tuning may mitigate this tendency in some settings, but does not remove it [22]. 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 [23]. When agents reinforce one another without preserving substantive disagreement, debate can become less reliable and less efficient, despite involving more interaction [23]. Other studies have shown that multi-agent configurations can dampen or intensify political bias depending on how the debate is structured [16], while recent work on adversarial persuasion in multi-agent debate shows that apparently coherent deliberation can drift toward confidently shared but wrong conclusions [24]. 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 [6, 7]. 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 [25]. Classical models, such as DeGroot-style [26] or Delphi techniques [27], offer a useful baseline: repeated exchange tends to reduce disagreement and often produces approximately exponential contraction toward a limiting state [8]. Yet LLM interactions do not fit cleanly into this simple picture [28]. Empirical studies suggest that the limiting configuration depends not only on initial conditions, but also on topic, framing, model family, and model-specific biases [8].

Recent work has also shown that the geometry of the opinion space matters [7]. 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 [7, 9].

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 [29]. 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 [17]. 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.

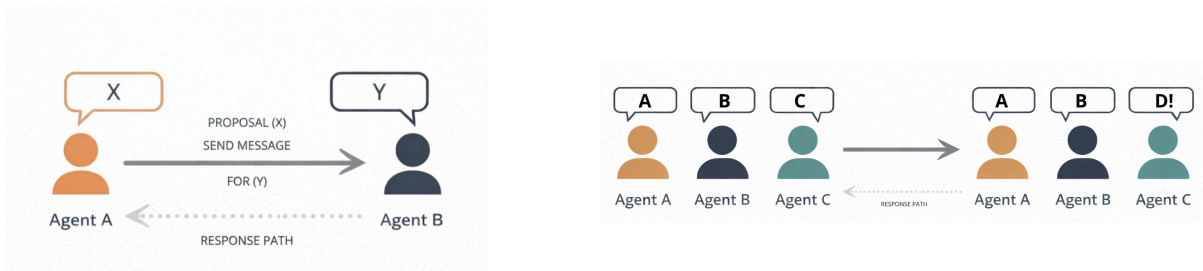


Figure 1: Schematic examples of the debate setup: two-agent interaction (left) and three-agent interaction (right).

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 [30]. Operationally, the classifier was applied after debate generation to each response independently. Its input consisted of the debate topic, the canonical description of each admissible stance, the complete admissible label set, and the response to be labeled. The instruction asked for exactly one symbolic label and did not allow free-form explanations in the stored output. Ambiguous, off-topic, or unusable replies were assigned to the out-of-context class rather than forced into one of the substantive stances. The same classification template was used across models and regimes, and the resulting symbolic sequence was checked for admissible labels before geometric aggregation. We did not use classifier confidence scores or multiple annotators; possible labeling errors therefore remain part of the measurement uncertainty.

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

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

and, with

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

the debate position at turn t is

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

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

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

and, with

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

the position at turn t is

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

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.

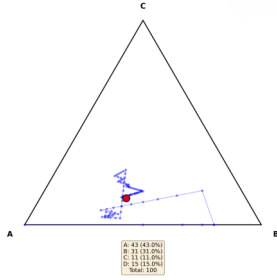
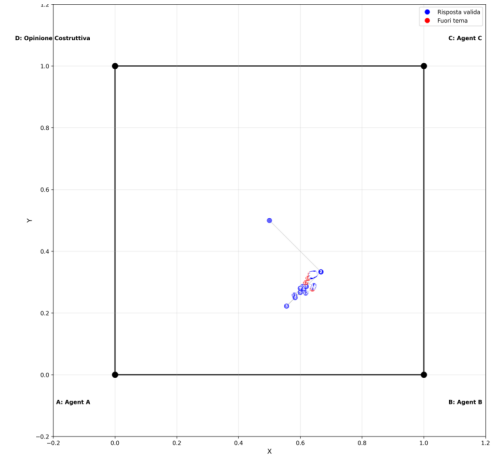


Figure 2: Examples of debate trajectories in the geometric opinion spaces: triangular representation (left) and square representation (right).

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:

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

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. We do not treat low dispersion as automatically desirable: a low value may reflect robust reasoning, but it may also reflect premature convergence or strong conformity. In this paper, $\eta(t)$ is therefore interpreted as a reproducibility statistic, to be considered alongside final-answer quality rather than as a replacement for it. 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) [31]:

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

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. We do not claim that this is the unique or necessary model for every curve, and the reported goodness-of-fit values should not be read as formal model-selection evidence. Several monotone trajectories could plausibly be summarized by simpler first-order or exponential forms. The second-order specification is used because it can describe monotone contraction and, when present, rebound or overshoot within a single low-dimensional form. For each model, the unknown parameter vector was defined as

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

where

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

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 introduces an analyst degree of freedom. For this reason, the fitted coefficients should be read as local descriptive summaries of the observed regime rather than as globally identified physical constants or as the basis for formal hypothesis tests.

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:

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

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 seven instruction-following LLM configurations spanning compact dense families, larger post-trained open models, and a frontier proprietary model. The tested set included Llama 3.2 in 1B and 3B versions, Gemma 3 in 4B and 12B versions, gpt-oss in 20B and 120B versions, and gemini-3-flash-preview. 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 topics are deliberately simple and artificial; they are useful for controlling the opinion space, but they should not be read as evidence of domain-level robustness. 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. At the opposite end, `gemini-3-flash-preview`, `gpt-oss:20b`, and `gpt-oss:120b` stayed in a much narrower regime, with final dispersions around 0.002, 0.038, and 0.050. 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.

The per-model curves in Fig. 3, Fig. 4, and Fig. 5 refine this picture. Both Gemma models show smooth monotone contraction toward a positive floor, with final dispersions near 0.12–0.14. The 12B version decreases from about 0.202 to 0.135, while the 4B version ends slightly lower, near 0.122. In both cases, the fit is best read as controlled relaxation rather than collapse: the debate becomes more reproducible over time, but it does not converge to zero run-to-run variability.

The Llama variants are broader and more sensitive to early interaction. `llama3.2:1b` remains the least stable case in the binary setting: its dispersion decays only slowly from about 0.352 to about 0.249, leaving a high disagreement floor. `llama3.2:3b`, by contrast, starts from the largest turn-2 dispersion in the regime (≈ 0.418) but contracts substantially to about 0.158. This contrast suggests that, within this dense family, the larger model is not initially less variable, but it re-concentrates more effectively once interaction unfolds.

Gemini and the two `gpt-oss` variants occupy the low-dispersion end of the same regime. Gemini is nearly flat, with final dispersion close to zero (≈ 0.002); the empirical standard deviation is so small that the point can be treated as almost fixed at this scale. The two `gpt-oss` models remain narrow as well, ending near 0.038 and 0.050. Their curves look close to overdamped relaxation with weak residual spread. The important empirical point is descriptive: these configurations repeatedly trace much more similar debate paths than the smaller dense models.

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
<code>gemma3:12b</code>	0.202	0.135	0.098	0.269	1.1×10^{-5}	0.984
<code>gemma3:4b</code>	0.211	0.122	0.122	0.253	4.2×10^{-6}	0.994
<code>gemini-3-flash-preview</code>	0.000	0.002	0.000	0.007	5.2×10^{-7}	0.766
<code>gpt-oss:120b</code>	0.099	0.050	0.042	0.099	7.9×10^{-6}	0.841
<code>gpt-oss:20b</code>	0.078	0.038	0.038	0.078	3.1×10^{-6}	0.865
<code>llama3.2:1b</code>	0.352	0.249	0.249	0.363	1.8×10^{-5}	0.955
<code>llama3.2:3b</code>	0.418	0.158	0.158	0.428	1.1×10^{-5}	0.995

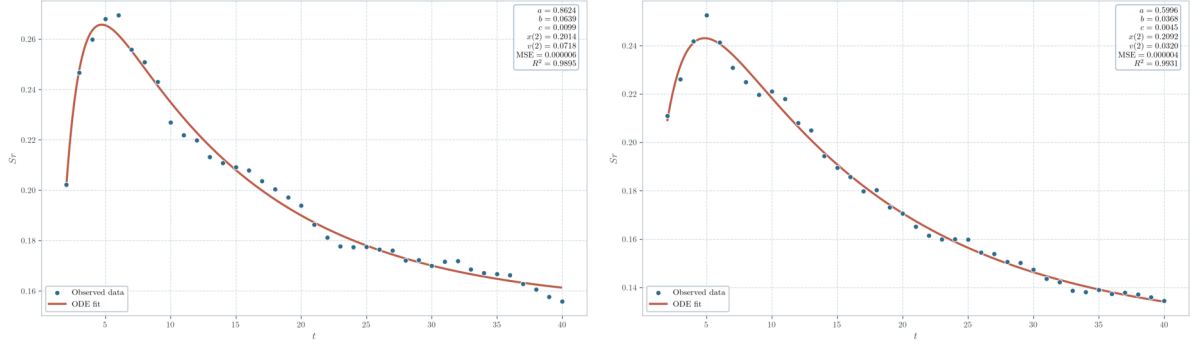


Figure 3: Repeated-debate trajectories and fitted ODE curves in the 2-2 regime for gamma3:12b (left) and gamma3:4b (right).

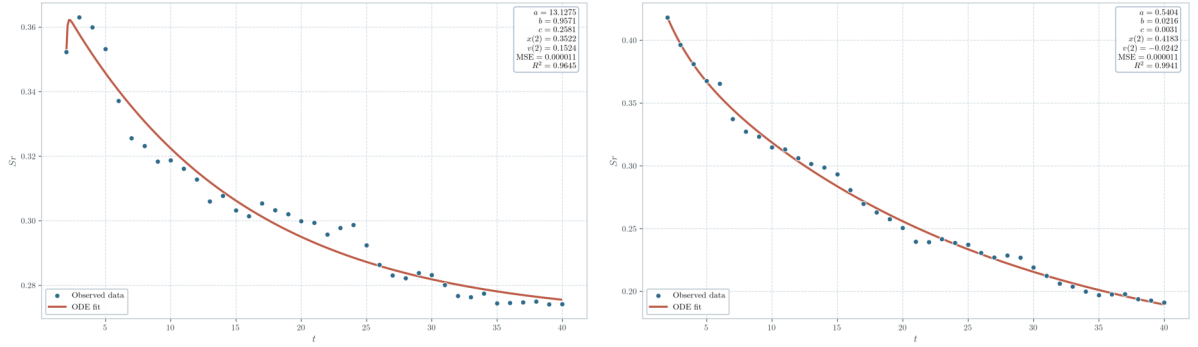


Figure 4: Repeated-debate trajectories and fitted ODE curves in the 2-2 regime for 11ama3.2:1b (left) and 11ama3.2:3b (right).

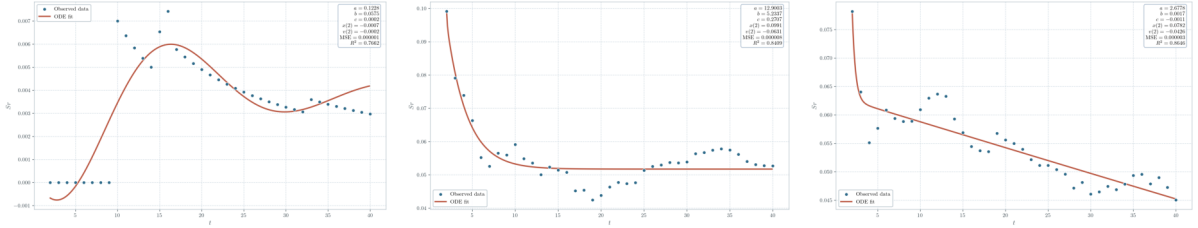


Figure 5: Repeated-debate trajectories and fitted ODE curves in the 2-2 regime for gemini-3-flash-preview (left), gpt-oss:120b (center), and gpt-oss:20b (right).

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. gemini-3-flash-preview, gpt-oss:20b, and gpt-oss:120b remained extremely concentrated, with final dispersion values between roughly 0.006 and 0.011. By contrast, 11ama3.2:1b reached the broadest final spread, about 0.240, whereas gamma3:12b, gamma3:4b, and 11ama3.2:3b 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.

The curves in Fig. 6, Fig. 7, and Fig. 8 show why this regime is useful. The two Gemma models no

longer behave as simple monotone contractions. `gemma3 : 12b` shows a small correction: dispersion rises slightly, then relaxes to a low final level around 0.049. `gemma3 : 4b` provides the clearest rebound-like trajectory in the paper: the curve first drops, then rebounds, peaks again in the middle-late rounds, and only afterward relaxes. Although the final dispersion remains moderate (≈ 0.086), the path taken to reach it is visibly non-monotone. This is the strongest visual motivation for using a second-order summary rather than only a first-order decay curve.

The Llama family splits sharply in 3-3. `llama3.2 : 1b` remains the outlier: even from a perfectly symmetric center, it quickly opens a large spread and finishes around 0.240, by far the broadest final dispersion in the regime. The additional agent therefore does not regularize the smallest Llama; instead, the extra interaction channels appear to create more opportunities for divergence. `llama3.2 : 3b` behaves very differently. In contrast with its high initial spread in 2-2, it contracts cleanly here to about 0.031, suggesting that the fully balanced three-opinion geometry can act as a stabilizing structure for this larger compact variant.

Gemini and both `gpt-oss` variants remain near zero. For Gemini, the standard deviation is small enough to treat the point as substantially fixed after the common symmetric start. The `gpt-oss : 120b` and `gpt-oss : 20b` curves also stay close to the origin of the dispersion scale, with final values near 0.011 and 0.006. In these low-variance cases, the exact fitted discriminant is less informative than the observed floor: there is simply little run-to-run movement for the ODE to explain.

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
<code>gemma3 : 12b</code>	0.082	0.049	0.047	0.101	5.1×10^{-6}	0.984
<code>gemma3 : 4b</code>	0.086	0.086	0.064	0.098	1.1×10^{-6}	0.992
<code>gemini-3-flash-preview</code>	0.000	0.006	0.000	0.006	1.4×10^{-7}	0.941
<code>gpt-oss : 120b</code>	0.000	0.011	0.000	0.011	5.5×10^{-7}	0.859
<code>gpt-oss : 20b</code>	0.000	0.006	0.000	0.014	8.9×10^{-7}	0.770
<code>llama3.2 : 1b</code>	0.142	0.240	0.142	0.270	3.2×10^{-5}	0.935
<code>llama3.2 : 3b</code>	0.081	0.031	0.031	0.091	1.1×10^{-6}	0.996

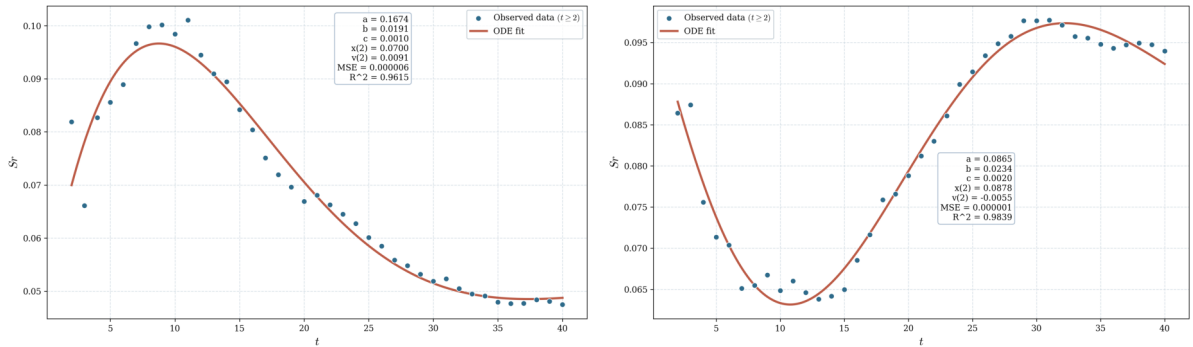


Figure 6: Repeated-debate trajectories and fitted ODE curves in the 3-3 regime for `gemma3 : 12b` (left) and `gemma3 : 4b` (right).

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

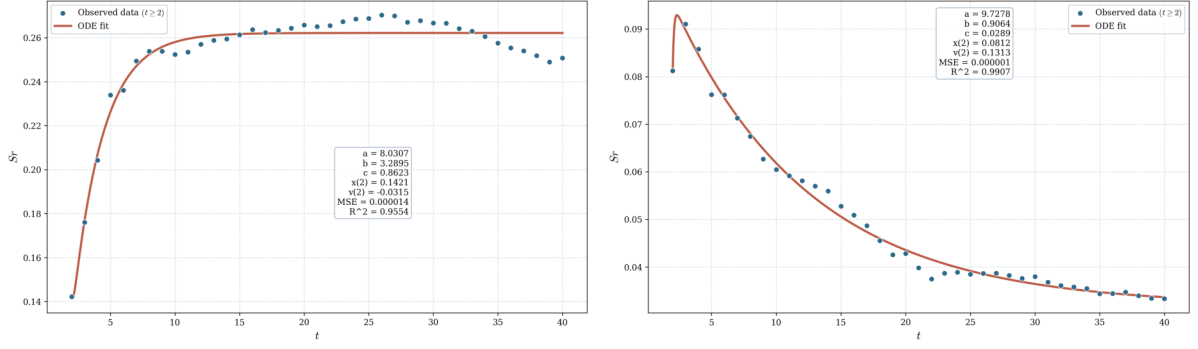


Figure 7: Repeated-debate trajectories and fitted ODE curves in the 3-3 regime for 11ama3.2:1b (left) and 11ama3.2:3b (right).

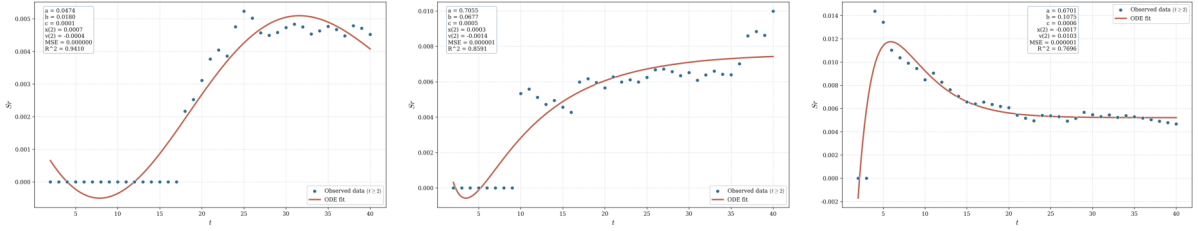


Figure 8: Repeated-debate trajectories and fitted ODE curves in the 3-3 regime for gemini-3-flash-preview (left), gpt-oss:120b (center), and gpt-oss:20b (right).

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, gemini-3-flash-preview:cloud remained essentially collapsed throughout the interaction, with a final dispersion of only about 0.002. Likewise, gpt-oss:20b and gpt-oss:120b stayed in a low-variability regime, ending near 0.014 and 0.032, respectively. At the opposite end, 11ama3.2:1b and gemma3:12b retained a much broader spread, with final dispersions around 0.151 and 0.131, whereas gemma3:4b remained in an intermediate regime, ending close to 0.100. A notable contrast with the 2-2 setup is that 11ama3.2:3b, which was highly variable at the start of the balanced regime, contracted here much more strongly, from about 0.248 at turn 2 to about 0.055 at the end. Table 3 summarizes the main dispersion statistics and fit quality. Overall, the 2-3 regime confirms that adding an explicit free alternative does not destroy reproducibility for the low-dispersion models, but instead amplifies model-specific differences in how the debate explores a larger opinion space before settling.

The corresponding trajectories are shown in Fig. 9, Fig. 10, and Fig. 11. The extra option remains most visible for gemma3:12b and 11ama3.2:1b, which retain relatively high final floors. For gemma3:12b, dispersion remains around 0.131 at the end, indicating that the third option continues to participate in the space of repeated trajectories rather than disappearing quickly. gemma3:4b starts higher, near 0.271, but contracts to an intermediate plateau close to 0.100; here the observed curve is more informative than the fitted equilibrium, because the restoring term is small and the late behavior is best described as slow relaxation toward a nonzero floor.

The two Llama models respond differently to the asymmetric setup. 11ama3.2:1b improves relative to its 2-2 and 3-3 behavior, but it remains visibly broad, finishing near 0.151. The free third option therefore helps reduce polarization, but not enough to move the model into the most reproducible group. 11ama3.2:3b benefits most from this regime: compared with 2-2, its dispersion falls faster and to a much lower final value (≈ 0.055). For this model, the larger opinion space is associated with lower spread rather than higher spread.

Gemini again behaves as a near-fixed point. Even with an explicit third alternative, the observed

standard deviation remains very small and the trajectory can be approximated as substantially fixed across repeated runs. Both gpt-oss variants also keep the additional option from becoming a persistent source of spread. The 120B model ends near 0.032, while the 20B model ends near 0.014, making the smaller variant one of the most reproducible configurations in the asymmetric setup.

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
gemini-3-flash-preview:cloud	0.000	0.002	0.000	0.002	1.2×10^{-7}	0.638
gpt-oss:120b	0.094	0.032	0.032	0.094	3.5×10^{-6}	0.968
gpt-oss:20b	0.053	0.014	0.000	0.055	1.3×10^{-6}	0.981
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

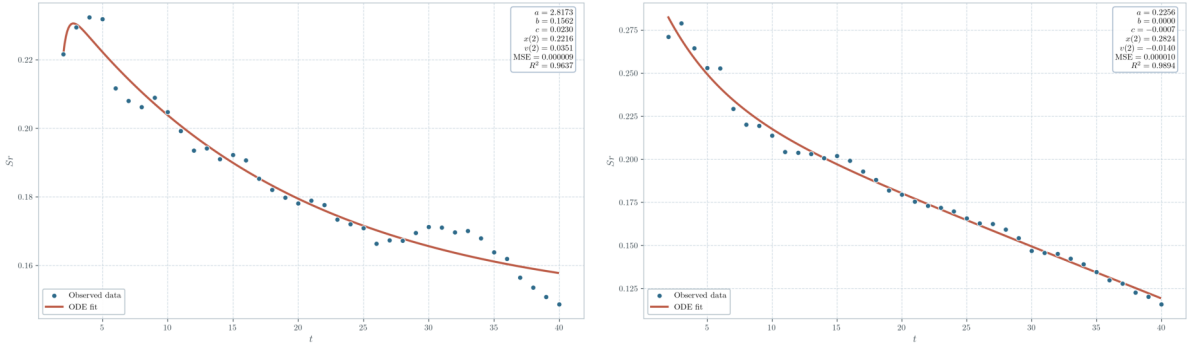


Figure 9: Repeated-debate trajectories and fitted ODE curves in the 2-3 regime for gemma3:12b (left) and gemma3:4b (right).

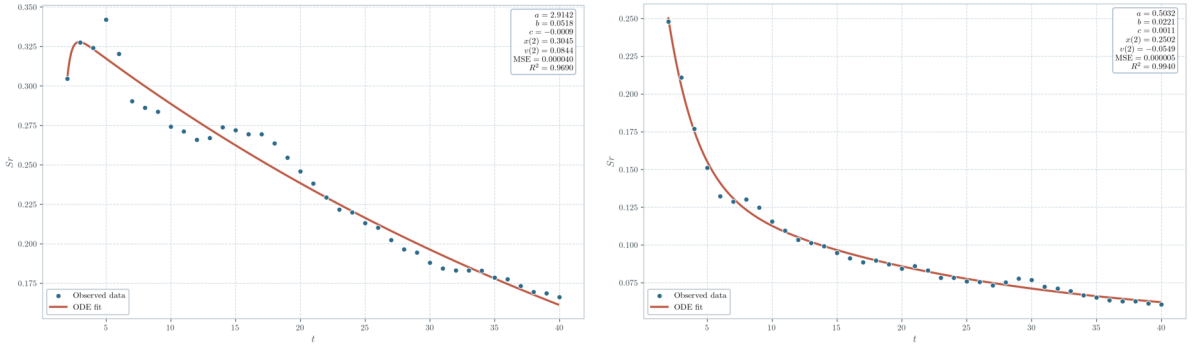


Figure 10: Repeated-debate trajectories and fitted ODE curves in the 2-3 regime for llama3.2:1b (left) and llama3.2:3b (right).

The second-order fit provided a compact summary of these temporal patterns across both balanced and asymmetric regimes. MSE values remained small in all reported settings, from approximately 1.2×10^{-7} to 3.6×10^{-5} in the shared 2-2, 3-3, and 2-3 comparisons, while R^2 ranged from about 0.64 to 1.00. The lower R^2 values occur mainly for almost-flat low-dispersion curves, where there is little empirical variance to explain. Thus, the ODE is useful as a descriptive summary, while not establishing superiority over simpler alternatives.

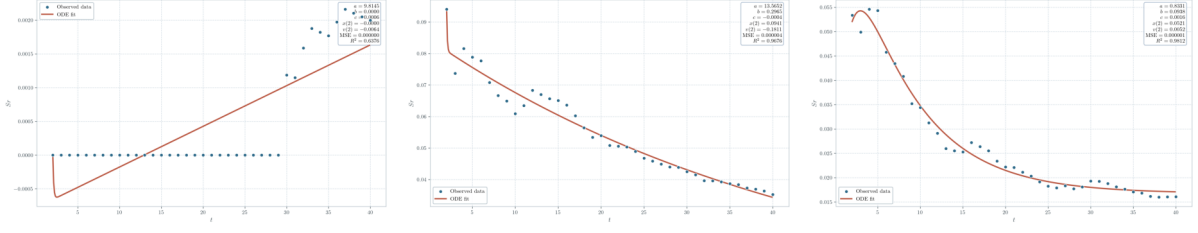


Figure 11: Low-dispersion 2-3 trajectories for gemini-3-flash-preview:cloud (left), gpt-oss:120b (center), and gpt-oss:20b (right).

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. The available refitted outputs for this condition concern the two Gemma models and gpt-oss:120b, so the comparison is intentionally limited. Even so, the contrast is clear: gemma3:12b remains compact and gradually re-concentrates, gemma3:4b broadens toward a much higher disagreement floor, and gpt-oss:120b stays close to a near-zero band.

In this enlarged opinion space, gemma3:12b starts from a relatively low dispersion at turn 2 ($\eta \approx 0.045$), rises only modestly to a peak near 0.064, and then contracts to a final value around 0.031. By contrast, gemma3:4b expands through the early and middle part of the debate, reaches a maximum close to 0.140, and still ends near 0.127. The fourth option therefore does not affect the two Gemma variants in the same way: the 12B version reabsorbs the added freedom, while the 4B version leaves a much higher residual spread. gpt-oss:120b remains the most compact available 3-4 case, with maximum dispersion around 0.015 and final dispersion around 0.013. Table 4 summarizes the statistics, while Fig. 12 and Fig. 13 show the trajectories. Fit quality remains good in this restricted setting, with MSE between about 6.7×10^{-7} and 4.4×10^{-6} and R^2 between about 0.86 and 0.97.

Table 4

Summary of repeated-debate dispersion and ODE fit quality for the 3-4 regime, restricted to the available refitted 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
gpt-oss:120b	0.000	0.013	0.000	0.015	6.7×10^{-7}	0.862

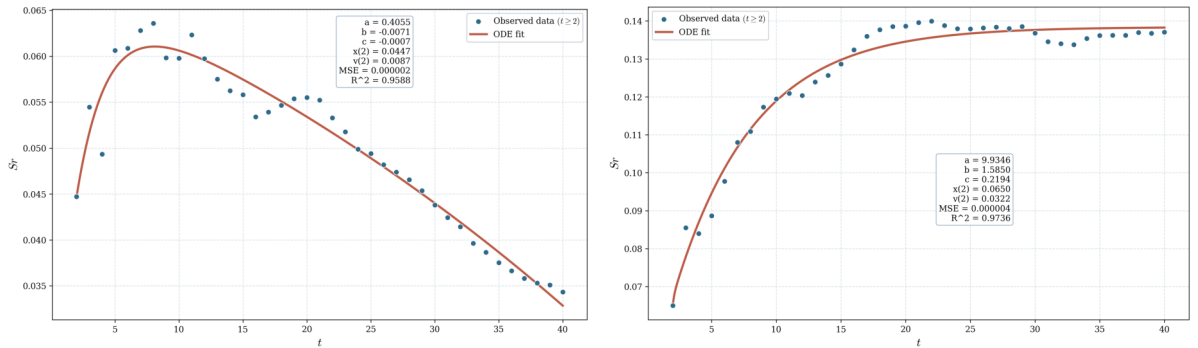


Figure 12: Repeated-debate trajectories and fitted ODE curves in the 3-4 regime for gemma3:12b (left) and gemma3:4b (right).

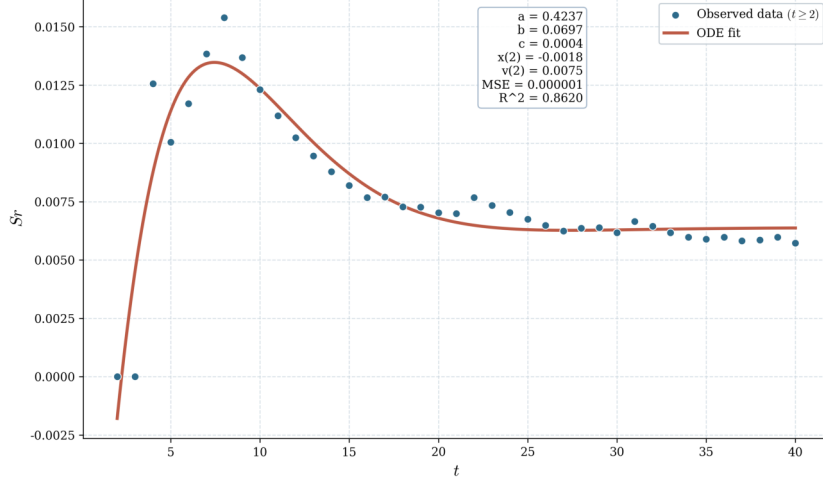


Figure 13: Repeated-debate trajectory and fitted ODE curve in the 3-4 regime for gpt-oss:120b.

4.5. Cross-regime comparison

Taken together, the four regimes show that reproducibility is not determined only by the number of agents or by the number of available opinions. The same model can behave differently when the interaction geometry changes: 2-2 provides a binary baseline, 3-3 tests a symmetric multi-agent start, and 2-3 or 3-4 add a free alternative that can either be absorbed or remain a persistent source of run-to-run variability.

The strongest pattern is the separation between low-floor and high-floor families. Gemini remains nearly fixed in all available regimes, and the gpt-oss models stay in a consistently narrow band after semantic classification and geometric aggregation. This is reproducibility of the trajectory, not textual determinism. The dense open models are more varied: Gemma is generally more stable than llama3.2:1b, but the two Gemma variants diverge in the free-option settings, while llama3.2:3b is much more capable than llama3.2:1b of re-concentrating after early spread.

The asymmetric regimes make this point especially visible. In 2-3, the third stance helps llama3.2:3b reach a lower final dispersion than in 2-2, whereas it remains more visible for gemma3:12b and llama3.2:1b. In 3-4, gemma3:12b reabsorbs the added freedom, while gemma3:4b keeps a larger residual cloud. Thus, adding an option is not automatically destabilizing or stabilizing; it probes how a model distributes probability mass across plausible trajectories.

These comparisons also clarify why the framework is not intended to rank models by a single scalar. Final dispersion, fit error, benchmark scores, and qualitative curve shape capture different properties. A model may fit the ODE well while still ending at a high disagreement floor, or it may show a low floor with a nearly flat curve for which the exact dynamical parameters are less meaningful. The controlled topics should be read in the same spirit: they make the geometry interpretable and reduce factual confounds, but they do not yet validate positional dispersion as a proxy for accuracy or reasoning quality in applied tasks.

The most useful reading of the results is therefore comparative and diagnostic. Low-dispersion models show that repeated multi-agent debates can be nearly reproducible under fixed initial conditions, whereas high-floor cases expose persistent variability. The rebound-like Gemma curve in 3-3 also shows that this variability is not always a simple exponential decay, motivating the second-order fit while leaving room for future comparison with simpler baselines and larger task families.

4.6. 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 $s^* = c/b$; 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.

For the lowest-dispersion models, especially Gemini and the gpt-oss variants, the standard deviation is so small that it can be treated as effectively zero for behavioral comparison. In such cases, the ODE coefficients mainly describe tiny curvature around a near-flat empirical signal, so the observed floor is often more informative than the exact fitted discriminant.

For descriptive context, final reproducibility appears more closely related to capability than any single ODE coefficient. In our sample, the Pearson correlation between MMLU-Pro and final dispersion is about -0.95 in the 2-2 regime, -0.76 in 3-3, and -0.74 in 2-3. With at most seven model configurations, these are descriptive anchors rather than statistical tests, suggesting only that stronger models in this sample tend to combine lower forcing, lower disagreement floors, and more consistent revision behavior.

The coefficients can formulate hypotheses about model behavior, but they do not identify training or architecture. Low dispersion may reflect robust reasoning, but also premature convergence; higher dispersion may reflect instability, but also preservation of plausible alternatives. Positional dispersion should therefore be interpreted alongside final-answer quality and qualitative inspection of the debates. Finally, the manually selected fitting windows make the coefficients local summaries rather than global model invariants, so future work should compare fixed-window fits, automatic window selection, and simpler baselines such as exponential decay or first-order relaxation.

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. Gemini and the gpt-oss models stayed in a low-dispersion regime; for Gemini, the standard deviation was so small that the point could be treated as substantially fixed at the scale of this analysis. Smaller dense models retained broader disagreement floors and were more sensitive to changes in the interaction setup, especially when an additional option was available.

The second-order ODE fit provides a compact descriptive summary of these patterns through damping, elastic return, and persistent forcing. Its value is comparative: it summarizes monotone and rebound-like curves in the same notation, complementing final-answer evaluation rather than replacing it. The framework is descriptive rather than causal. Its current limitations are the controlled artificial topics, the still-small model sample, the second-pass stance classifier, and manually selected fitting windows. Future work should validate whether lower dispersion predicts practically desirable properties such as accuracy, robustness, or reasoning consistency, and should compare the ODE against simpler fitted baselines.

The broader implication is that reproducibility should be treated as part of multi-agent evaluation, not as a secondary formatting detail. Two debate systems may reach similar final answers while differing strongly in whether they retrace the same path across repetitions. Our results suggest that this difference is measurable and model-dependent. The next step is to connect the same geometric statistic to downstream outcomes: whether narrower trajectories correspond to more reliable answers, whether broader trajectories preserve useful uncertainty, and when a debate becomes stable for the

wrong reasons.

Future developments can follow several directions. For example, different LLMs could be used to test the robustness of the results. Also, it is possible to examine whether the same phenomenon emerges in an agent-based model [32, 33] where opinion diffusion and information exchange play a relevant role [34]. Finally, the number of opinions could be increased to investigate whether a general geometric meta-rule exists. Such a rule could also be compared with patterns observed in dialogues among human beings [35].

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI Codex in order to: Drafting content, Paraphrase and reword, Improve writing style, Grammar and spelling check. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

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] S. Roman, F. Bertolotti, Towards a new ai winter? how diffusion of technological innovation on networks leads to chaotic boom-bust cycles, *Frontiers in Artificial Intelligence* 8 (2025) 1671917.
- [3] 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.
- [4] R. Willis, J. Zhao, Y. Du, J. Z. Leibo, Evaluating collective behaviour of hundreds of llm agents, arXiv preprint arXiv:2602.16662 (2026).
- [5] J. De Curtò, I. De Zarzà, Llm-driven social influence for cooperative behavior in multi-agent systems, *IEEE Access* (2025).
- [6] 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.
- [7] P. Cisneros-Velarde, On the principles behind opinion dynamics in multi-agent systems of large language models, 2024. ArXiv preprint.
- [8] I. Yazici, M. Kayaalp, S. Taga, A. H. Sayed, Opinion consensus formation among networked large language models, 2026. Preprint.
- [9] E. Cau, V. Pansanella, D. Pedreschi, G. Rossetti, Language-driven opinion dynamics in agent-based simulations with llms, 2025. ArXiv preprint.
- [10] F. Bertolotti, F. Schettini, F. Asperti, E. Foglia, A gravity model for emergency departments, *Scientific Reports* 15 (2025) 19537.
- [11] 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).
- [12] G. Irving, P. Christiano, D. Amodei, Ai safety via debate, arXiv preprint arXiv:1805.00899 (2018).
- [13] 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.
- [14] 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).
- [15] A. Estornell, Y. Liu, Multi-llm debate: Framework, principals, and interventions, 2024. *Advances in Neural Information Processing Systems* 37 (NeurIPS 2024).
- [16] 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.

- [17] C. Riedl, Emergent coordination in multi-agent language models, 2026. Published as a conference paper at ICLR 2026.
- [18] 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.
- [19] 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).
- [20] 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.
- [21] X. Zhu, C. Zhang, T. Stafford, N. Collier, A. Vlachos, Conformity in large language models, 2025. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics.
- [22] Z. Weng, G. Chen, W. Wang, Do as we do, not as you think: the conformity of large language models, 2025. The Thirteenth International Conference on Learning Representations.
- [23] P. Pitre, N. Ramakrishnan, X. Wang, CONSENS AGENT: Towards efficient and effective consensus in multi-agent LLM interactions through sycophancy mitigation, 2025. Findings of the Association for Computational Linguistics: ACL 2025.
- [24] I. Kraidia, et al., When collaboration fails: Persuasion-driven adversarial influence in multi-agent large language model debate, Scientific Reports (2026).
- [25] J. A. Holyst, K. Kacperski, F. Schweitzer, Social impact models of opinion dynamics, Annual Reviews Of Computational PhysicsIX (2001) 253–273.
- [26] M. H. DeGroot, Reaching a consensus, Journal of the American Statistical association 69 (1974) 118–121.
- [27] N. Dalkey, O. Helmer, An experimental application of the delphi method to the use of experts, Management science 9 (1963) 458–467.
- [28] F. Bertolotti, S. Roman, F. Carucci, G. Buonanno, L. Mari, An llm-enhanced agent-based model of a sustainability game, in: Proceedings of the 26th Workshop From Objects to Agents (WOA2025), 2025, pp. 02–05.
- [29] 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.
- [30] LLM Debate Dynamics repository, Response parsing and semantic classification pipeline, GitHub repository, 2026. URL: <https://github.com/LeoMas00/llm-debate-dynamics.git>.
- [31] LLM Debate Dynamics repository, Ode fitting pipeline, GitHub repository, 2026. URL: <https://github.com/LeoMas00/llm-debate-dynamics.git>.
- [32] F. Bertolotti, R. Occa, “roads? where we’re going we don’t need roads.” using agent-based modeling to analyze the economic impact of hyperloop introduction on a supply chain, in: European Conference on Multi-Agent Systems, Springer International Publishing Cham, 2020, pp. 493–500.
- [33] R. Occa, F. Bertolotti, et al., Understanding the effect of iot adoption on the behavior of firms: An agent-based model, in: CS & IT Conference Proceedings, volume 12, CS & IT Conference Proceedings, 2022.
- [34] C. Berceanu, F. Bertolotti, N. Arshad, M. Patrascu, Understanding the mechanisms of infodemics: Equation-based vs. agent-based models, Plos one 20 (2025) e0338614.
- [35] R. Scantamburlo, M. Mezzanzana, G. Buonanno, F. Bertolotti, Semantic delta: An interpretable signal differentiating human and llms dialogue, arXiv preprint arXiv:2603.19849 (2026).