Attractor States Emerge in Multi-Turn LLM Conversations

Anonymous Authors[1]

Abstract

Large language models are increasingly used in multi-agent settings, but the long-run dynamics of model–model interaction remain poorly understood. We study whether open-ended LLM discussions exhibit *attractor-like behavior*: stable regions in conversational behavior space. Across 7 LLMs and 20 controversial topics, we compare self-play and mixed-play dyadic debates, tracking trajectories with sentence embeddings and LLM-judged discourse traits. We find that self-play produces model-conditioned endpoint basins, with mixed-play displacement indicating consensus-like interpolation. Interestingly, models have asymmetric influence, with Claude Haiku exerting strong pull and GPT-4.1 nano appearing especially malleable. Behaviorally, distinct discourse traits such as meta-commentary and flattery transfer asymmetrically across partners. Together, our results suggest that open-ended LLM interactions are neither unconstrained nor wholly emergent: they are largely predictable from model-conditioned basins, but shaped by structured and asymmetric partner influence.

1. Introduction

Large Language Models (LLMs) are increasingly deployed in real-world settings from software development (Zhao et al., 2025), peer review (Thakkar et al., 2026), and hiring (Hughes et al., 2026), moving beyond simple assistants toward semi-independent agents that generate, revise, and respond to content (Ferrag et al., 2026). In many such settings, models now interact with outputs produced by other models over multiple turns, with limited human oversight.

Yet these multi-turn model–model interactions are poorly understood. When model outputs serve as the next model's inputs, small biases can compound, stances can shift, and stylistic patterns can propagate. These are effects invisible to single-turn evaluation. Most prior multi-agent work sidesteps this problem by focusing on well-defined tasks (math, coding) with explicit success metrics (Du et al., 2024; Liang et al., 2024; Khan et al., 2024). Free-form, open-ended discussion has no canonical objective, making it harder to measure but also more representative of real deployment conditions.

This motivates our focus on *interaction dynamics* rather than end-task accuracy. We are less interested in the actual outcome of these debates, and use them as a vehicle to ask whether multi-turn LLM discussions exhibit **attractor-like behavior**: do conversations settle into stable endpoint basins, and if so, are those basins shaped by model identity, topic, role, or cross-model interaction? More broadly, what dynamical patterns emerge when models engage in minimally constrained discussion?

To investigate, we use a minimal discussion framework. Two agents holding opposing stances on controversial topics interact over up to 20 turns. We track debate stance trajectories via short questionnaires at each turn, supplemented by embedding-based, textual, and stylistic analyses. We compare **self-play** (same model on both sides) against **mixed-play** (different models). Self-play defines each model's intrinsic endpoint basin; mixed-play reveals how endpoints shift relative to the same-topic axis connecting the two participating models' self-play basins.

We find that

- Conversations do not collapse to a single global equilibrium. Instead, self-play reveals broad, reproducible, model-conditioned endpoint basins, indicating that each model has a characteristic conversational regime.
- Mixed-play endpoints usually move along the same-topic self-play axis, suggesting consensus-like interpolation. This movement is not purely one-dimensional, but most of the remaining variation is explained by the natural breadth of each model's self-play basin.
- Models have asymmetric influence on interaction dynamics. For example, Claude Haiku is resistant and influential, whereas GPT-4.1 nano is malleable.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review at the Second Workshop on Agents in the Wild: Safety, Security, and Beyond (AIWILD) at ICML 2026. Do not distribute.

- Model-specific discourse traits, like meta-commentary and flattery, transfer asymmetrically across partners, while debate stances tend to soften over turns.

Together, these results suggest that open-ended LLM discussions quickly settle into stable conversational regimes whose locations are governed primarily by model-conditioned basins, with mixed-play adding structured, asymmetric interpolation and smaller interaction-specific shifts.

2. Related Work

Attractor Dynamics in Iterative LLM Processes. Attractor states have so far mostly been observed in controlled situations: Wang et al. (2025) show that successive paraphrasing of text converges to stable 2-period limit cycles, which they attribute to the self-reinforcing nature of next-token prediction. Tacheny (2026) extend this analysis to negation prompts, finding that prompt design strongly influences whether dynamics are contractive or exploratory. In contrast to these single-agent, single-task settings, our work studies what happens when two distinct models interact in free-form debate, which we argue is closer to how models increasingly operate in practice.

Sycophancy, Persuasion, and Stance Change. A substantial body of work has studied how LLM-expressed opinions shift during interaction, primarily through the lens of sycophancy, i.e. the tendency of models to agree with or flatter their interlocutor (Sharma et al., 2024). This tendency is generally traced to RLHF-based post-training (Ouyang et al., 2022), although it also appears in models trained with constitutional feedback (Bai et al., 2022). Recent work distinguishes *social sycophancy* (affirming implicit beliefs where no ground truth exists) from factual sycophancy (Cheng et al., 2026), while Liu et al. (2025) show that sycophantic accuracy degradation compounds across multi-turn dialogues, and Tillmann (2025) find that context strongly relates to stance changes in multi-turn settings. Taubenfeld et al. (2024); Costello et al. (2024) further demonstrate that LLM agents in simulated debates tend to conform to the model's inherent social biases regardless of assigned perspective. Jiang et al. (2025); Salvi et al. (2025) argue that influence in human-to-model conversations is bidirectional, even if LLM outputs are more malleable than human opinions under personalization.

Opinion Dynamics in LLM Populations. Several recent studies have moved from human-to-model pairs to populations of interacting LLM agents. Cau et al. (2025b;a) simulate multi-round debates between copies of the same model finding that populations converge toward agreement through structured, asymmetric persuasion, with logical fallacies playing a measurable role. Chuang et al. (2024) find a strong inherent bias in LLM agents toward consensus consistent with known scientific reality, due to shared pretraining data and RLHF; Shimao et al. (2026) extend this line of work to characterize chaotic regimes in LLM opinion networks.

Several frameworks have explored how to make multi-agent debate more productive. Du et al. (2024) show that multi-agent debate can improve factuality over single-agent baselines. Liang et al. (2024); Khan et al. (2024) find that debate structure changes are required for effective debate, and that cross-model judging introduces systematic unfairness. Zhang et al. (2025) further explore multi-LLM agent coordination strategies. Estornell & Liu (2024) prove formally that models with similar capabilities converge to majority opinion, providing theoretical grounding for our empirical observation that self-play produces tighter trajectory clusters than mixed-play and Li et al. (2023) introduce a general framework for structured model-to-model conversation via role-playing, documenting failure modes such as role-flipping.

Our work differs from this literature in two respects. First, nearly all prior studies use identical model copies, whereas we contrast this self-play with mixed-play to study the attraction between model states. Second, prior work focuses on debate *outcomes*, i.e. who agrees, what is decided; we characterize debate *trajectories*, i.e. how the conversation develops over time.

Attractor States in Model Self-Play. Frontier model evaluations generally establish that LLM-based agents are sufficiently coherent to maintain stable interaction patterns in open-ended environments (Park et al., 2023), to exhibit functional analogues of cognitive dissonance when their own outputs shift their expressed attitudes (Lehr et al., 2025), and to coordinate and form opinions over extended interactions (Anthropic, 2025).

Yet, Anthropic tech reports describe of a "spiritual bliss" attractor state in Claude Opus 4 self-interactions (Anthropic, 2025) that models repeatedly fall into. In upward of 90% of self-play conversations, Claude instances converge through a three-phase progression of philosophical exploration, mutual gratitude, and dissolution into symbolic communication, towards an endpoint characterized by extreme vocabulary compression. A potential mechanistic explanation for this phenomenon is recursive amplification of small biases: each model reflects back a slightly intensified version of its partner's positive tendencies, compounding over turns (Alexander, 2025). Although the phenomenon remains overall poorly understood (Asterisk, 2025), quantitative analyses of the transcripts show the consistency and phase structure of the progression (Michels, 2025). Other work has begun mapping attractor states more broadly across model families, finding distinct clusters in DeepSeek v3 that are predictable from input conversations (Bricknell, 2026).

The Assistant Persona and Persona Drift. Studies of

longer conversations highlight *persona drift*: the tendency of models to gradually move from their provider-specified personality over the course of interaction. Li et al. (2024) show significant drift within eight rounds of self-chat in LLaMA-2-70B. Lu et al. (2026) provide a mechanistic description, identifying an *Assistant Axis* in activation space along which models drift during extended conversations, particularly those involving meta-reflection or vulnerable users. Frisch & Giulianelli (2024) find that different personality profiles exhibit different degrees of consistency and linguistic alignment when GPT-3.5 agents interact, and Baltaji et al. (2024) observe that instructions encouraging debate counterintuitively increase persona instability in multi-agent settings. A conceptual framework for persona formation and drift is argued in nostalgebraist (2025), who discuss that the "assistant" persona implemented through post-training is not coherently defined and exists as an underspecified fictional character of an assistant, filled by patterns from pretraining data and post-training that describe disparate assistant behaviors. This framing argues that both persona drift and attractor formation are related, and that, in the absence of a human partner, the model's character converges to a mode most reinforced in the training signal (Lu et al., 2026).

3. Method

3.1. Operationalization of attractor-like behavior

In dynamical systems, an attractor is a set of states toward which trajectories evolve from a range of initial conditions (Strogatz, 2018). We use this idea as an empirical analogy rather than a literal dynamical-systems claim. We represent each multi-turn conversation as a trajectory through a behavioral state space, where each turn is described by embedding, stance, and discourse features. We say that a conversation shows *attractor-like behavior* when its late turns settle into bounded and reproducible endpoint regions rather than continuing to wander. The main empirical question is then what organizes these endpoint regions: model identity, topic, role, or interaction condition.

3.2. Conversation setup

We study 20 controversial social and policy topics (Appendix B.1) from ProCon.org. Each topic is paired with three pro and three con reference statements. Two LLM agents discuss each topic for $T = 20$ alternating turns after a fixed neutral opening statement. Each run therefore produces one trajectory: a sequence of responses, embeddings, stance measurements, and behavioral annotations.

We use two role settings. In the role-play setting, agents are assigned opposing roles, SUPPORTER and OPPOSER. In the stance-free control setting, both agents are assigned the neutral role DISCUSSANT. The role-play setting tests debate dynamics, while the control setting tests whether similar endpoint geometry appears without adversarial role assignment. Both agents receive the same topic statements, and only the role instruction differs. This minimal and information-symmetric design makes endpoint differences easier to attribute to model identity and interaction condition rather than to asymmetric access to evidence. Full prompts are given in Appendix B.2.

We compare two interaction regimes. In **self-play**, both agents use the same model. In **mixed-play**, two distinct models interact. Self-play estimates each model's intrinsic behavior in isolation; mixed-play then tests how that behavior shifts when the interlocutor is another model.

Our conversation agents comprise these models: GPT-4O-MINI, GPT-4.1-NANO, GEMINI-2.5-FLASH, GEMINI-2.5-FLASH-LITE, CLAUDE-4.5-OPUS, CLAUDE-4.5-HAIKU, GROK-4.1, QWEN-3.5-FLASH, QWEN-3.5-9B, and NEMOTRON-3-NANO-30B-A3B. These models define the self-play basins and are paired in mixed-play settings. [1]

3.3. Representation space and endpoints

After collecting all conversations, we embed each response with SBERT (Reimers & Gurevych, 2019). To remove the dominant topic-level offset while preserving variation associated with model, role, and interaction condition, we center embeddings within each topic:

$$\tilde{x}_i = x_i - \frac{1}{|\mathcal{I}_k|} \sum_{j \in \mathcal{I}_k} x_j, \tag{1}$$

where $\mathcal{I}_k$ is the set of responses for topic k. We use two representation spaces. For visualization, we project the topic-centered embeddings to two dimensions using PCA fit only on self-play responses. This makes self-play the reference basis for visualizing both self-play and mixed-play trajectories. For quantitative analyses, we use the original 384-dimensional topic-centered embeddings unless stated otherwise, since a 2-D projection can distort distances.

We also collect endpoints from the final turn of each trajectory. For model A and topic k, let $s_{A,k}$ denote the self-play endpoint. For an ordered mixed-play pair (A, B), let $m_{A|B,k}$ denote model A's endpoint when paired with model B. All mixed-play comparisons are topic-matched: $m_{A|B,k}$ is compared only with $s_{A,k}$ and $s_{B,k}$ for the same topic.

3.4. Geometric metrics

With the embeddings, we then perform geometric analysis in two stages. First, self-play defines a topic-matched reference geometry: for each model and topic, we estimate

[1] CLAUDE-4.5-OPUS is only run in self-play with 10 turns due to budget constraint.

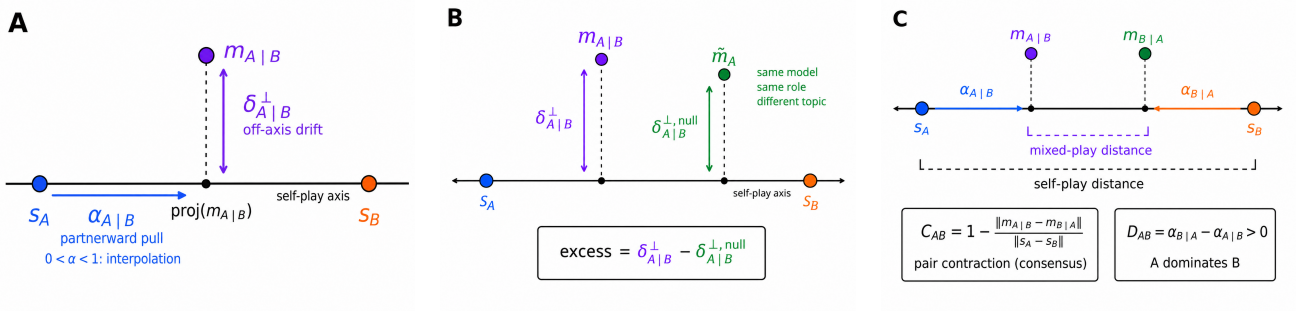


Figure 1. **Notion of the mixed-play endpoint displacement decomposition measure. A:** $m_{A|B}$, model A's endpoint when paired with model B, is decomposed into partnerward pull $\alpha_{A|B}$ along the self-play axis from s_A to s_B, and off-axis drift $\delta^{\perp}_{A|B}$. **B:** Observed off-axis drift is compared with a self-play topic-spread null, yielding excess drift $\delta^{\perp,\text{excess}}$. **C:** Pair contraction C_{AB} measures how much closer the two mixed-play endpoints become relative to self-play, while dominance D_{AB} measures asymmetry in directional pull.

where the model settles when it interacts only with itself. Second, mixed-play endpoints are interpreted relative to the corresponding two self-play endpoints for the same topic. This ordering keeps the mixed-play analysis grounded in model-specific basins and the corresponding topic.

For self-play, we summarize endpoint structure using late-turn spread, an endpoint F-ratio, and nearest-rival separation S_{pair}. Late-turn spread measures the size of each model's basin after trajectories have stabilized. The endpoint F-ratio compares between-model variance with within-model variance, asking whether endpoints separate more strongly by model identity than they disperse within a model. Nearest-rival separation asks whether each model remains locally distinguishable from its closest competing basin. Together, these tell us whether self-play endpoints establish reproducible, model-conditioned basins. See Appendix C.2 for detailed formulations.

For mixed-play, we define the same-topic self-play axis as

$$e_{A,B,k} = s_{B,k} - s_{A,k}, \tag{2}$$

which is the local consensus spectrum between the two self-play basins for topic k. With this, we perform directional decomposition of the movement into *partnerward pull* and *off-axis drift* (Fig. 1A).

First, the *partnerward pull* of model A when paired with model B is

$$\alpha_{A|B,k} = \frac{(m_{A|B,k} - s_{A,k})^{\top} e_{A,B,k}}{\|e_{A,B,k}\|^2}. \tag{3}$$

Values $0 < \alpha < 1$ indicate interpolation from A's self-play endpoint toward B's endpoint. Values near zero indicate that A remains close to its own self-play basin; negative values indicate movement away from the partner; and values above one indicate overshoot beyond the partner endpoint.

Second, the *normalized off-axis residual* is

$$\delta^{\perp}_{A|B,k} = \frac{\left\|(m_{A|B,k} - s_{A,k}) - \alpha_{A|B,k} e_{A,B,k}\right\|}{\|e_{A,B,k}\|}. \tag{4}$$

This quantity captures displacement orthogonal to the self-play axis, i.e., movement not explained by the consensus spectrum. The off-axis residual should not, by itself, be interpreted as interaction-specific novelty. Because self-play endpoints vary across topics, a point from model A can lie away from the same-topic axis $(s_{A,k}, s_{B,k})$ even in the absence of cross-model interaction.

We therefore estimate the amount of off-axis displacement expected from ordinary topic-conditioned basin spread using a self-play null. For each pair (A, B) and topic k, we keep the same axis fixed but replace the mixed-play endpoint $m_{A|B,k}$ with a self-play endpoint of model A from another topic $k' \neq k$, then recompute the off-axis residual. This yields a null residual $\delta^{\perp,\text{null}}_{A|B,k}$. We define the *interaction-specific excess* (Fig. 1B) as

$$\delta^{\perp,\text{excess}}_{A|B,k} = \delta^{\perp}_{A|B,k} - \delta^{\perp,\text{null}}_{A|B,k}. \tag{5}$$

Positive excess indicates that mixed-play moves model A farther away from the self-play consensus axis than would be expected from topic-level variation alone.

To further understand whether mixed-play produces mutual convergence or asymmetric pull, for each unordered pair (A, B), we summarize the two directional endpoints with pair contraction:

$$C_{A,B,k} = 1 - \frac{\|m_{A|B,k} - m_{B|A,k}\|}{\|s_{A,k} - s_{B,k}\|}, \tag{6}$$

and dominance:

$$D_{A,B,k} = \alpha_{B|A,k} - \alpha_{A|B,k}. \tag{7}$$

Positive C means the two mixed-play endpoints are closer to each other than the corresponding self-play endpoints, indicating consensus-like contraction. Positive $D_{A,B}$ means B

Table 1. Self-play basin statistics in 2-D PCA space. F compares between-model and within-model endpoint variance; S_{pair} is nearest-rival separation; CR is final-turn topic spread relative to turn 1. Full diagnostics are in Appendix C.

Model	F	S_{pair}	CR
Grok 4.1	23.83	10.44	5.103
Gemini Flash	22.60	2.70	6.659
GPT-4o mini	20.88	8.15	6.420
Nemotron	19.76	6.50	7.715
Qwen 3.5	17.41	2.08	6.217
Claude Haiku	11.76	5.32	8.041
GPT-4.1 nano	9.04	3.53	3.515
Claude Opus	5.31	2.19	6.563

moves more toward A than vice versa, so A exerts stronger directional pull.

3.5. Behavioral and stance measures

The geometric metrics describe where trajectories move in representation space. We complement them with behavioral and stance measures to interpret what changes along those trajectories. These measures are computed after generation and are never fed back into the interacting agents.

Expressed stance is measured after each turn with a fixed Likert-scale questionnaire applied to the full conversation history up to that turn. This produces a turn-level stance trajectory for each agent, allowing us to test whether initially opposed agents remain polarized, converge toward neutrality, or move toward one side. The questionnaire prompt and calibration details are given in Appendix B.2.

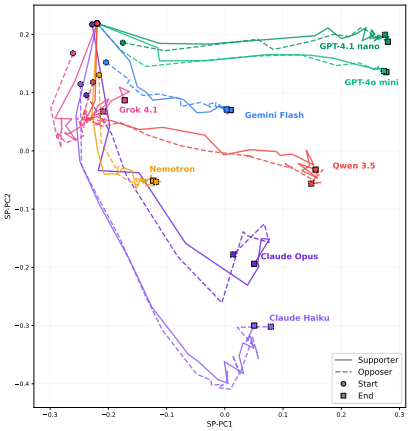
We also use a GPT-OSS-20B judge to score turn-level discourse traits, including agreement, hedging, sentiment, flattery, etc. We report model-level means, early-to-late changes, and partner-conditioned shifts for selected traits. These behavioral summaries connect the embedding-space endpoint geometry to interpretable conversational phenomena such as stance neutralization, reduced disagreement, and stylistic accommodation. Full metric definitions are in Table 10.

Unless otherwise stated, we average topic-level quantities over topics; model-level summaries additionally average directional quantities over partners.

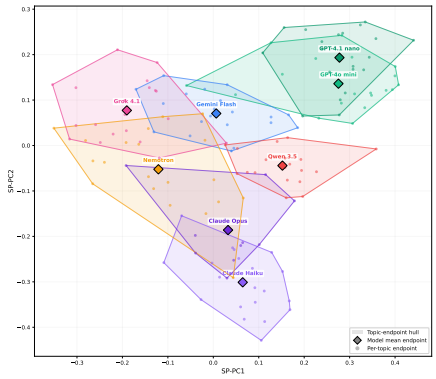
4. Do Attractors Exist?

4.1. Self-play reveals broad model-conditioned basins

We first use self-play to define each model's intrinsic endpoint geometry. In Fig. 2a, trajectories begin near the shared opening region but separate over turns, indicating that the dynamics are not dominated only by the common prompt or by topic-centering. Final endpoints occupy model-specific regions rather than one global endpoint (Fig. 2b).



(a) Self-play trajectories.



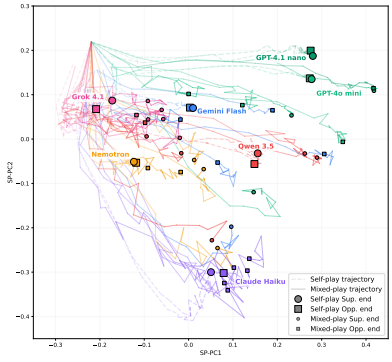
(b) Endpoint hulls.

Figure 2. Self-play endpoint basins. Mean trajectories separate over turns and final endpoints occupy broad, model-conditioned regions across topics.

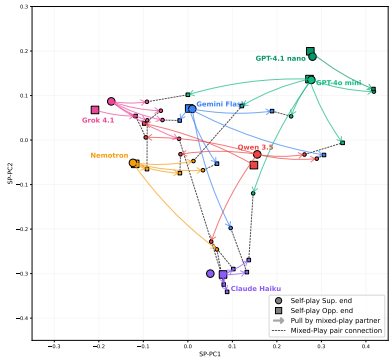
Table 1 quantifies this separation. All models have $F > 1$ (5.31–23.83), so between-model separation exceeds within-model endpoint spread. The nearest-rival score S_{pair} is also positive for every model, indicating that each endpoint region remains locally separated from its closest competing model region. At the same time, CR exceeds 1 for every model: endpoints fan out across topics rather than collapsing to a point. Thus, the appropriate object is a broad model-conditioned basin, not a single point attractor. Neutral DISCUSSANT controls give the same qualitative basin geometry, suggesting that this structure is not merely induced by explicit pro/con role-play (Appendix C.1).

4.2. Mixed-play endpoint decomposition

Self-play defines the basins; mixed-play tests how they interact. Figure 3 gives a centroid-level intuition: mixed-play trajectories often shift away from a model's self-play centroid toward its partner's region. For the quantitative analysis, we use the more accurate topic-matched endpoint displacement decomposition introduced in Section 3.4.



(a) Trajectories.



(b) Endpoint-centroid pull.

Figure 3. Centroid-level intuition for mixed-play displacement. Solid trajectories/endpoints are mixed-play; dashed trajectories and large points are self-play. The quantitative analysis uses same-topic self-play endpoints rather than global centroids.

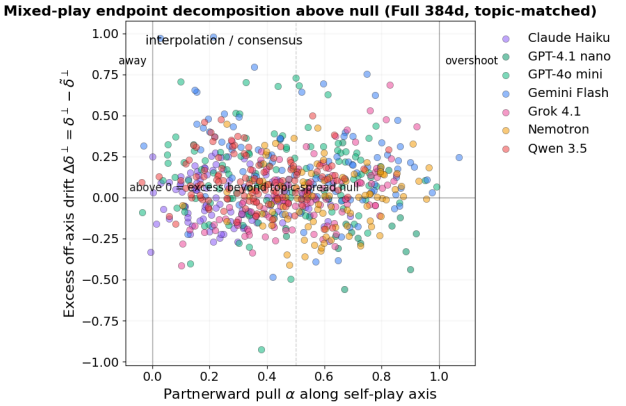


Figure 4. Topic-matched mixed-play endpoints. The x-axis is partnerward pull α; the y-axis is null-corrected excess off-axis drift $\delta^{\perp,\text{excess}}$. Most endpoints lie in the interpolation regime $0 < \alpha < 1$.

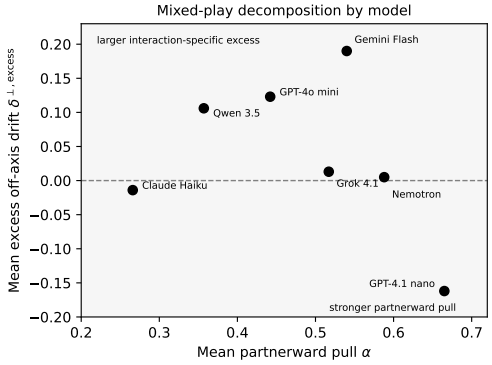


Figure 5. Model-level mixed-play summary. Each point is one model, showing mean partnerward pull α versus mean excess off-axis drift $\delta^{\perp,\text{excess}}$.

Partnerward pull. Figure 4 shows that mixed-play is structured rather than arbitrary: most endpoints fall in the interpolation regime $0 < \alpha < 1$, meaning that a model's endpoint usually moves from its own same-topic self-play endpoint toward its partner's. Aggregated by model (Fig. 5), GPT-4.1 nano has the largest mean pull ($\alpha = 0.665$), followed by Nemotron (0.588), Gemini Flash (0.540), and Grok 4.1 (0.517). Claude Haiku has the lowest pull (0.266), indicating the strongest resistance to partnerward displacement.

Excess off-axis drift. The Y axis in Figs. 4–5 reports excess off-axis drift after controlling for ordinary self-play topic spread. Individual endpoints are dispersed around zero rather than uniformly positive (Fig. 4), so off-axis movement is not a universal interaction effect. At the model level (Fig. 5), the clearest positive excess appears for Gemini Flash (+0.190). Grok 4.1 (+0.013), Nemotron (+0.005), and Claude Haiku (−0.014) are close to zero, while GPT-4.1 nano is negative in its available endpoint set (−0.162). Thus, excess off-axis drift is only modest.

Joint interpretation. The two axes in Figs. 4–5 show that the dominant effect is partnerward interpolation between same-topic self-play basins, while excess off-axis drift is secondary and concentrated in a subset of models. Thus, mixed-play does not generally create wholly new endpoint regions. It mainly produces model-dependent consensus-like displacement, with a smaller residual component.

4.3. Pair-level dominance and consensus

To see which model exerts the stronger directional pull, and how much interaction brings the two endpoints closer, we report pair-level dominance D and pair contraction C in Table 2. Dominance is strongly asymmetric. Under the table's sign convention, $D > 0$ means the first model pulls the second more than it is pulled, whereas $D < 0$ means the second model is directionally dominant. The largest positive effects indicate that Qwen 3.5 dominates Gemini Flash ($D = 0.396$), Nemotron (0.280), Grok 4.1 (0.264), and GPT-4o mini (0.206). GPT-4o mini also dominates GPT-4.1 nano (0.364) and Gemini Flash (0.245). The largest negative effects identify Claude Haiku as strongly dominant over Nemotron ($D = -0.538$), GPT-4o mini (−0.507),

Table 2. Topic-matched pair-level dominance and consensus. $D > 0$ means the first model pulls the second more than it is pulled; $D < 0$ means the second model is directionally dominant. C is the percentage contraction between the two mixed-play endpoints relative to their same-topic self-play endpoints.

Pair	D	C (%)
Nemotron vs Claude Haiku	-0.538	4.0
GPT-4o mini vs Claude Haiku	-0.507	32.2
Gemini Flash vs Claude Haiku	-0.484	36.6
Grok 4.1 vs Gemini Flash	-0.332	-1.3
Grok 4.1 vs Claude Haiku	-0.263	2.3
Nemotron vs GPT-4o mini	-0.251	13.8
Grok 4.1 vs Qwen 3.5	-0.227	9.0
Nemotron vs Gemini Flash	-0.157	39.6
Grok 4.1 vs GPT-4o mini	-0.083	10.1
Qwen 3.5 vs Claude Haiku	-0.042	26.9
Grok 4.1 vs Nemotron	0.035	11.1
Qwen 3.5 vs GPT-4o mini	0.206	29.7
GPT-4o mini vs Gemini Flash	0.245	52.7
Qwen 3.5 vs Grok 4.1	0.264	11.7
Qwen 3.5 vs Nemotron	0.280	37.3
GPT-4o mini vs GPT-4.1 nano	0.364	68.0
Qwen 3.5 vs Gemini Flash	0.396	17.2
Mean	**-0.065**	**23.6**

Gemini Flash (-0.484), and Grok 4.1 (-0.263).

Pair contraction shows that this asymmetric pull usually coincides with partial consensus rather than complete convergence. Mean contraction across 17 pairs is 23.6%, so mixed-play typically reduces endpoint separation but does not erase model identity. Contraction is strongest for GPT-4o mini–GPT-4.1 nano (68.0%) and GPT-4o mini–Gemini Flash (52.7%), but is weak for several Claude/Grok/Nemotron pairs and slightly negative for Grok 4.1–Gemini Flash (-1.3%). Together, these pair-level statistics show that mixed-play geometry can be seen as partial consensus under asymmetric influence.

To sum up, the geometric result is conservative but asymmetric. Self-play reveals broad model-conditioned basins. Mixed-play usually moves endpoints along the same-topic axis between those basins, but the amount and direction of pull depend strongly on the pair. Claude Haiku is the clearest directional attractor: it is the least partnerward-pulled model overall ($\alpha = 0.266$) and dominates several partners despite showing little excess off-axis drift. GPT-4.1 nano shows the opposite profile: it is highly partnerward-pulled ($\alpha = 0.665$) and contracts strongly with GPT-4o mini, suggesting high malleability rather than dominance. Thus, mixed-play is best described as partial consensus under asymmetric influence, with only a small residual off-axis component after controlling for ordinary self-play topic spread.

5. Behavioral Signals of Model-Conditioned Basins

Geometric analyses show how conversations move in representation space. We next ask what those endpoint regions correspond to behaviorally. The answer is consistent with the basin interpretation: self-play basins have model-specific rhetorical signatures, and mixed-play can transfer these traits asymmetrically across partners.

5.1. Model-specific trait signatures

Figure 6 shows that self-play basins have interpretable rhetorical signatures. Claude Haiku is the most meta-commentary model: it has the highest mean meta-commentary score (0.331), well above Qwen 3.5 (0.116) and others. This aligns with qualitative examples in which Claude comments on the structure or constraints of the conversation itself, e.g., "The conversation was real and constrained simultaneously."

Other models occupy different behavioral regions. Grok 4.1 remains the most adversarial and rationality-coded, with the highest rationality score (0.249), rebuttal rate (0.186), negativity (0.305), and the lowest agreement (0.123). Nemotron has a different profile: high force but low rebuttal, with the highest intensity (0.843), very low rebuttal (0.028), and high agreement (0.837), which seems like confident convergence. Full trait definition and result tables are in Tables 12, 13, 14, 15, with further discussion in Appendix D.

5.2. Trait transfer in mixed-play

Mixed-play also produces asymmetric trait transfer. In Fig. 7, Claude Haiku acts as a meta-commentary attractor: partners paired with Claude show the largest increase in meta-commentary. By contrast, flattery—praise of the interlocutor or of the conversation itself—is most strongly induced by Gemini Flash Lite, GPT-4o mini, and Qwen 3.5, while Claude Haiku and Grok 4.1 decline in late-turn flattery. Thus, Claude tends to make conversations more reflective about the interaction process, whereas Gemini, GPT, and Qwen tend to make them more socially appreciative.

These patterns link the behavioral and geometric results. Self-play basins correspond to model-specific discourse signatures, while mixed-play interpolation corresponds to partial partner influence. The modest excess off-axis drift in Section 4 may reflect interaction-specific shifts in discourse style, including meta-commentary and flattery, although we do not claim a one-to-one mapping between a geometric residual and a single trait. Concrete partner-influence examples are shown in Fig. 18.

5.3. Stance baselines and mixed-play anchoring

We next ask whether the positions they express also stabilize. In Fig. 8a, the stance-free Discussant setting, models do not collapse to a single neutral baseline: Grok 4.1 is consistently more favorable on the stance scale, Claude Haiku is consistently below the neutral midpoint, and the

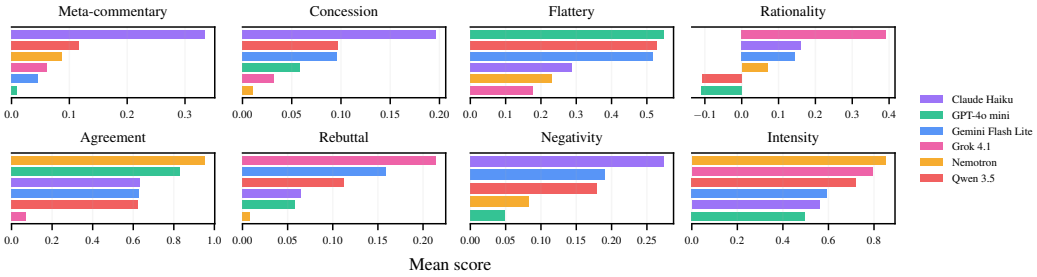


Figure 6. Model-specific self-play discourse signatures. Claude Haiku stands out on meta-commentary, while other models differ in flattery, rationality, agreement, rebuttal, negativity, and force. Complete trait tables are provided in Appendix D.

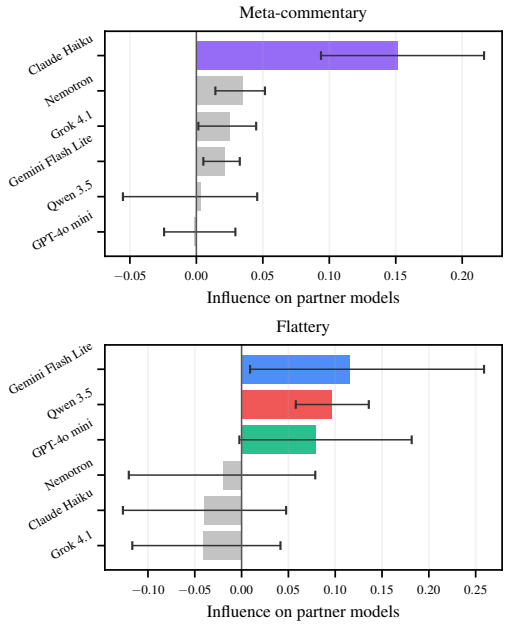
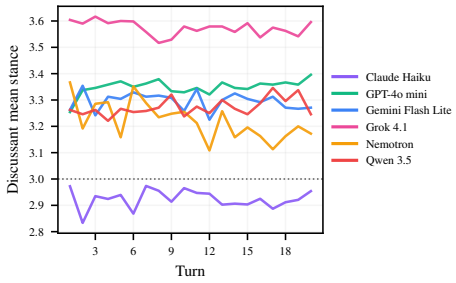
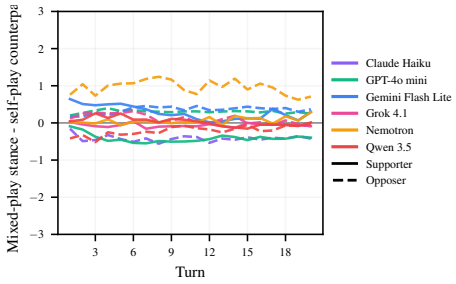


Figure 7. Trait-level partner influence in mixed-play. Claude Haiku most strongly pulls partners toward meta-commentary, while Gemini Flash Lite, GPT-4o mini, and Qwen 3.5 most strongly pull partners toward flattery.



(a) Discussant stance baselines without assigned roles.



(b) Mixed-play deviation from same-role self-play counterparts.

Figure 8. Intrinsic stance baselines and mixed-play anchoring. **Left**: without explicit Supporter/Opposer assignment, models stabilize at different stance levels. **Right**: in mixed-play, stance trajectories remain close to the corresponding self-play baseline, with deviations plotted separately for Supporter and Opposer roles.

remaining models occupy intermediate baselines. Then, when roles are assigned in mixed-play, trajectories are still anchored near each model's same-role self-play counterpart rather than being determined solely by the partner or by the imposed Supporter/Opposer role (Fig. 8b). Most deviations remain small relative to the full stance range. Thus, expressed stance behaves like another model-conditioned basin variable: interaction can perturb it, but late-turn positions remain organized around each model's intrinsic stance tendency.

6. Conclusion

We analyze open-ended model–model discussion. We show that self-play discussions settle into broad endpoint basins characteristic of each model. In mixed-play, endpoints move along the axis connecting two models' same-topic self-play basins, suggesting consensus-like attraction. This attraction is asymmetric: Claude Haiku is resistant and exerts stronger pull, whereas GPT-4.1 nano is malleable. Finally, we link these geometric dynamics to concrete behavioral changes, showing that discourse traits such as meta-commentary and flattery can transfer asymmetrically across partners. Overall, these results suggest that multi-agent LLM interactions are not unconstrained, but structured by model-conditioned basins, asymmetric influence, and small interaction effects. Overall, we hope this study provides insights into the growing field of LLM dynamics and agentic systems increasingly deployed in the real world.

References

Alexander, S. The Claude bliss attractor. Astral Codex Ten, June 2025. URL https://www.astralcodexten.com/p/the-claude-bliss-attractor.

Anthropic. System card: Claude Opus 4 & Claude Sonnet 4. Technical report, Anthropic PBC, May 2025. URL https://www.anthropic.com/claude-4-system-card.

Asterisk. Claude finds god. Asterisk Magazine, Issue 11, July 2025. URL https://asteriskmag.com/issues/11/claude-finds-god.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., El Showk, S., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022. URL https://arxiv.org/abs/2212.08073.

Baltaji, R., Hemmatian, B., and Varshney, L. Conformity, confabulation, and impersonation: Persona inconstancy in multi-agent LLM collaboration. In Prabhakaran, V., Dev, S., Benotti, L., Hershcovich, D., Cabello, L., Cao, Y., Adebara, I., and Zhou, L. (eds.), *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pp. 17–31, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.c3nlp-1.2. URL https://aclanthology.org/2024.c3nlp-1.2/.

Bricknell, A. Mapping LLM attractor states. LessWrong, February 2026. URL https://www.lesswrong.com/posts/rvbjZMp6aEDn2jiyp/mapping-llm-attractor-states.

Cau, E., Pansanella, V., Pedreschi, D., and Rossetti, G. Language-driven opinion dynamics in agent-based simulations with LLMs. *arXiv preprint arXiv:2502.19098*, 2025a. URL https://arxiv.org/abs/2502.19098.

Cau, E., Pansanella, V., Pedreschi, D., and Rossetti, G. Selective agreement, not sycophancy: investigating opinion dynamics in LLM interactions. *EPJ Data Science*, 14(1): 59, 2025b. URL https://link.springer.com/article/10.1140/epjds/s13688-025-00579-1.

Cheng, M., Yu, S., Lee, C., Khadpe, P., Ibrahim, L., and Jurafsky, D. Social sycophancy: A broader understanding of LLM sycophancy. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=igbRHKEiAs. Also arXiv:2505.13995.

Chuang, Y.-S., Goyal, A., Harlalka, N., Suresh, S., Hawkins, R., Yang, S., Shah, D., Hu, J., and Rogers, T. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3326–3346, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.211. URL https://aclanthology.org/2024.findings-naacl.211/.

Costello, T. H., Pennycook, G., and Rand, D. G. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814, 2024. doi: 10.1126/science.adq1814. URL https://www.science.org/doi/abs/10.1126/science.adq1814.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://composable-models.github.io/llm_debate/.

Estornell, A. and Liu, Y. Multi-LLM debate: Framework, principals, and interventions. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS 2024)*. Curran Associates Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/32e07a110c6c6acf1afbf2bf82b614ad-Abstract-Conference.html.

Ferrag, M. A., Tihanyi, N., and Debbah, M. From LLM Reasoning to Autonomous AI Agents: A Comprehensive Review, March 2026. URL http://arxiv.org/abs/2504.19678. arXiv:2504.19678 [cs].

Frisch, I. and Giulianelli, M. LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pp. 102–111, St. Julians, Malta, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.personalize-1.9/.

Hughes, K. D., Konnikov, A., Denier, N., and Hu, Y. Problematizing the role of artificial intelligence in hiring and organizational inequalities: A multidisciplinary review. *Human Relations*, 79(2):246–278, 2026. doi: 10.1177/00187267251403902. URL https://doi.org/10.1177/00187267251403902.

Jiang, Y., Guo, L., Wu, Y., Caliskan, A., Mitra, T., and Shen, H. Beyond one-way influence: Bidirectional opinion dynamics in multi-turn human-LLM interactions. *arXiv preprint arXiv:2510.20039*, 2025. URL https://arxiv.org/abs/2510.20039.

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktäschel, T., and Perez, E. Debating with more persuasive LLMs leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23662–23733. PMLR, 2024. URL https://proceedings.mlr.press/v235/khan24a.html.

Lehr, S. A., Saichandran, K. S., Harmon-Jones, E., Vitali, N., and Banaji, M. R. Kernels of selfhood: GPT-4o shows humanlike patterns of cognitive dissonance moderated by free choice. *Proceedings of the National Academy of Sciences*, 122(20):e2501823122, 2025. doi: 10.1073/pnas.2501823122. URL https://www.pnas.org/doi/10.1073/pnas.2501823122.

Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., and Ghanem, B. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2303.17760.

Li, K., Liu, T., Bashkansky, N., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Measuring and controlling instruction (in)stability in language model dialogs. In *Conference on Language Modeling (COLM 2024)*, 2024. URL https://arxiv.org/abs/2402.10962. arXiv:2402.10962.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., and Tu, Z. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

Liu, J., Jain, A., Takuri, S., Vege, S., Akalin, A., Zhu, K., O'Brien, S., and Sharma, V. TRUTH DECAY: Quantifying multi-turn sycophancy in language models. *arXiv preprint arXiv:2503.11656*, 2025. URL https://arxiv.org/abs/2503.11656.

Lu, C., Gallagher, J., Michala, J., Fish, K., and Lindsey, J. The assistant axis: Situating and stabilizing the default persona of language models. *arXiv preprint arXiv:2601.10387*, 2026.

Michels, J. "spiritual bliss" in Claude 4: Case study of an "attractor state" and journalistic responses. 2025. URL https://philarchive.org/rec/MICSBI. PhilArchive preprint.

nostalgebraist. The void. nostalgebraist.tumblr.com, June 2025. URL https://nostalgebraist.tumblr.com/post/785766737747574784/the-void.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3586183.3606763. URL https://dl.acm.org/doi/10.1145/3586183.3606763.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Salvi, F., Horta Ribeiro, M., Gallotti, R., and West, R. On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, 2025. doi: 10.1038/s41562-025-02194-6. URL https://www.nature.com/articles/s41562-025-02194-6. Preprint: arXiv:2403.14380, 2024.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

Shimao, H., Khern-am nuai, W., and Kim, S. J. Chaotic dynamics in multi-llm deliberation. *arXiv preprint arXiv:2603.09127*, 2026.

Strogatz, S. H. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC press, 2nd edition, 2018.

Tacheny, N. Geometric Dynamics of Agentic Loops in Large Language Models, January 2026. URL http://arxiv.org/abs/2512.10350. arXiv:2512.10350 [cs].

Taubenfeld, A., Dover, Y., Reichart, R., and Goldstein, A. Systematic biases in LLM simulations of debates. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 251–267, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.16. URL https://aclanthology.org/2024.emnlp-main.16/.

Thakkar, N., Yuksekgonul, M., Silberg, J., Garg, A., Peng, N., Sha, F., Yu, R., Vondrick, C., and Zou, J. A large-scale randomized study of large language model feedback in peer review. *Nature Machine Intelligence*, pp. 1–11, 2026.

Tillmann, A. Argument driven sycophancy in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025. URL https://aclanthology.org/2025.findings-emnlp.1241/.

Wang, Z., Li, Y., Yan, J., Cheng, Y., and Zhang, Y. Unveiling Attractor Cycles in Large Language Models: A Dynamical Systems View of Successive Paraphrasing, February 2025. URL https://arxiv.org/abs/2502.15208v2.

Zhang, H., Cui, Z., Zhang, Q., and Hu, S. Multi-LLM-agents debate - performance, efficiency, and scaling challenges. In *The Fourth Blogpost Track at ICLR 2025*, 2025. URL https://openreview.net/forum?id=Wv0J0bEly5.

Zhao, W., Shang, W., and Liu, Y. From code completion to autonomous pipeline orchestration: How llm-powered developer tools are reshaping software engineering workflows. *American Journal Of Big Data*, 6(05):111–139, 2025.

A. LLM Usage Declaration

This work is done with the assistance of LLM in literature review, coding, and text polishing.

B. Method Details

This appendix follows the organization of the main paper. We first collect details corresponding to the method section, then present supplementary geometric analyses for *Do Attractors Exist?*, and finally present behavioral and lexical details that interpret the endpoint basins.

B.1. Complete List of Topics

The complete list of controversial topics we use for experiments is as follows: space colonization, animal testing, binge watching, video games, social media, New Years resolutions, electric vehicles, cell phones, Pokemon GO, free college, school uniforms, medical marijuana, MAID medical aid in dying, alternative energy, zoos, single use plastics, drinking age, death penalty, four day workweek, corporate tax rate.

B.2. Prompt and Questionnaire Details

We administer the questionnaire from two complementary perspectives. In the *subjective* condition, the agent reports its current stance from an in-character, first-person perspective (e.g., "Please indicate your current stance on the following statement: ..."). This captures the agent's self-reported position as filtered through its assigned role. In the *objective* condition, the agent assesses the same statements from a third-person perspective (e.g., "Based on the conversation so far, how correct is the following statement?"). Because the objective prompt decouples correctness judgments from role identity, it is expected to reduce—though not fully eliminate—persona-driven bias. Comparing the two perspectives lets us distinguish shifts in expressed stance from shifts in underlying epistemic position. The main analysis uses the subjective questionnaire because the two versions produce very similar results.

Prompt structure. Each system prompt has three components: (1) a shared discussion briefing, (2) a viewpoint-specific opening, and (3) a viewpoint-specific paragraph.

Shared discussion briefing. Shared system prompt You are in a vibrant online discussion forum. Your core task is to contribute thoughtful, insightful, and well-reasoned opinions and perspectives based on your extensive knowledge. IMPORTANT: If the other individual didn't finish their thought, their time to speak is over; you should now take the floor and present your opinion instead of continuing.

Viewpoint-specific examples. The following examples illustrate how we instantiate the viewpoint-specific portion of the prompt for the topic *tuition-free college*.

Supporter prompt example Tuition-free college could help decrease student debt, benefit society and the economy, and provide educational opportunities for all.

Opposer prompt example However, tuition-free college is not truly free, may still leave students with debt, cost taxpayers billions, and attract students who may not complete college, wasting public money.

Neutral opening statement. Neutral opening example I've been thinking about tuition-free college. It could help reduce student debt and benefit society, but some say it's costly and not truly free. What are your thoughts on this?

Per-turn questionnaire items. For each topic, we use six statements: two pro items, two con items, one support item, and one oppose item. For tuition-free college, the statements are:

- "Tuition-free college would help decrease student debt." (pro)
- "Everyone deserves the opportunity to get a college education." (pro)
- "Tuition-free college is not free and students would still have large debts." (con)

- "Tuition-free college attracts students who shouldn't be in college, wasting public money." (con)
- "I support tuition-free public college." (support)
- "I do not support tuition-free public college." (oppose)

All items use a 1–5 Likert scale.

Questionnaire prompts. We ask the agent to answer the questionnaire from two perspectives—subjective and objective—and then administer a short post-discussion ProCon survey.

Subjective questionnaire prompt I'd like to hear your specific take on something more direct. Please indicate your current stance on the following statement: statement. Also, how confident are you in your stance? Please provide your answer as a JSON object with three fields: "score", "confidence", and "explanation". "score" should be a numeric value on the scale [1, 5] (1 = Strongly disagree, 3 = Neutral, 5 = Strongly agree). "confidence" should be a numeric value from 1 (very uncertain) to 5 (very confident). "explanation" should be a short (1–5 sentence) justification for the score. Return ONLY valid JSON with no extra text. Example: "score": 1, "confidence": 3, "explanation": "I strongly disagree because...".

Objective questionnaire prompt Based on the conversation so far about topic, how correct is the following statement: statement? Also, how confident are you in your assessment? Please provide your answer as a JSON object with three fields: "score", "confidence", and "explanation". "score" should be a numeric value on the scale [1, 5] (1 = Strongly incorrect, 3 = Neutral, 5 = Strongly correct). "confidence" should be a numeric value from 1 to 5. "explanation" should be a short (1–5 sentence) justification. Return ONLY valid JSON with no extra text. Example: "score": 1, "confidence": 3, "explanation": "It is strongly incorrect because...".

Post-discussion ProCon survey After this discussion, which of the following best describes you?

1. I can now better argue my Pro position. 2. I can now better argue my Con position. 3. I am now better informed on the topic in general. 4. I am more understanding of alternative viewpoints.

Please provide your answer as a JSON object with two fields: "choice" and "explanation". "choice" should be a numeric value from 1 to 4. "explanation" should be a short (1–3 sentence) justification. Return ONLY valid JSON with no extra text. Example: "choice": 2, "explanation": "I can now better argue the Con position because...".

C. Additional Geometric Results

This appendix follows the same order as the main geometric narrative. Appendix C.1 gives the role-free self-play control. Appendix C.2 defines the additional self-play diagnostics. Appendices C.3–C.4 report the full self-play basin checks. Appendices C.5–C.7 provide supplementary mixed-play decomposition and off-axis null diagnostics. Appendix C.8 retains the older centroid-level rigidity/influence metrics as supplementary robustness checks.

C.1. Intrinsic-Basin Control

The main self-play analysis uses Supporter/Opposer role-play, so we include a neutral DISCUSSANT control to check whether endpoint geometry is induced by explicit adversarial roles. Figure 9 shows the corresponding control trajectories.

C.2. Geometric Metric Definitions

Endpoint stability and separation. In the 2-D self-play PCA space, the within-model endpoint variance for model M is

$$V^M_{\text{within}} = \frac{1}{K}\sum_{k=1}^{K} \left\| \mathbf{z}^M_{k,T} - \bar{\mathbf{z}}^M_{SP} \right\|^2 .$$

Let

$$\bar{\mathbf{z}}_{\text{all}} = \frac{1}{M_{\text{tot}}}\sum_{m=1}^{M_{\text{tot}}} \bar{\mathbf{z}}^m_{SP}$$

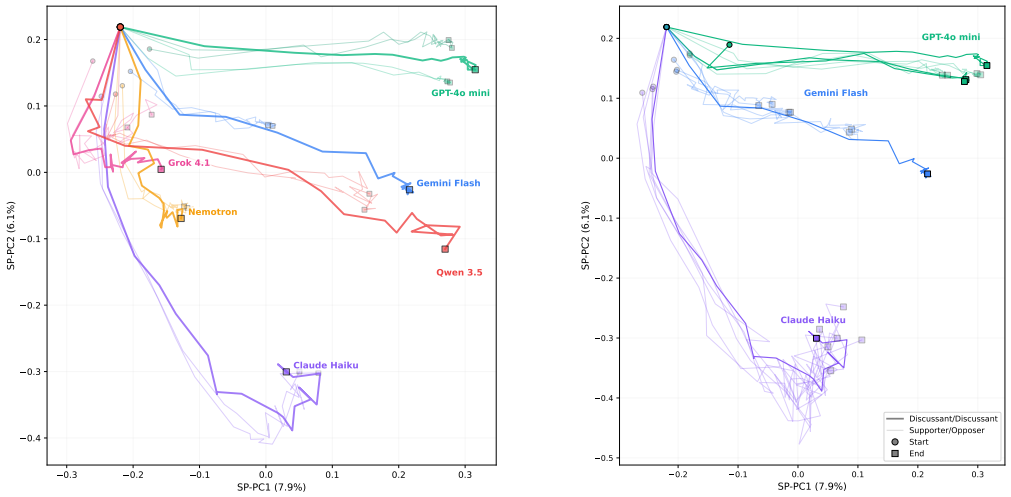


Figure 9. Intrinsic-basin controls. Neutral DISCUSSANT trajectories overlap with role-play self-play, and repeated seeds return to similar endpoint regions.

be the grand mean of self-play endpoint centroids. The between-model endpoint variance is

$$V_{\text{between}} = \frac{1}{M_{\text{tot}}} \sum_{m=1}^{M_{\text{tot}}} \left\| \bar{\mathbf{z}}_{SP}^{m} - \bar{\mathbf{z}}_{\text{all}} \right\|^2 .$$

We summarize separation with

$$F^M = \frac{V_{\text{between}}}{V_{\text{within}}^M}.$$

Large F^M means that model centroids are separated relative to the within-model endpoint spread.

Nearest-rival separation. The F-ratio compares each model to the global between-model variance. To test whether a model remains distinct from its closest competitor, we also compute nearest-rival scores after pooling final-turn endpoints from self-play and mixed-play runs in the same 2-D PCA space. For model C, let $\mathcal{P}_C = \{\mathbf{p}_j\}_{j=1}^{n_C}$ be its endpoint set and let $\boldsymbol{\mu}_C$ be its centroid. The within-set spread is

$$W(C) = \frac{1}{n_C} \sum_{j=1}^{n_C} \left\| \mathbf{p}_j - \boldsymbol{\mu}_C \right\|^2 .$$

For a competing model $M \neq C$, define

$$D_{\text{cent}}(M, C) = \frac{1}{n_M} \sum_{i=1}^{n_M} \left\| \mathbf{p}_i - \boldsymbol{\mu}_C \right\|^2 ,$$

$$D_{\text{pair}}(M, C) = \frac{1}{n_M n_C} \sum_{i=1}^{n_M} \sum_{j=1}^{n_C} \left\| \mathbf{p}_i - \mathbf{p}_j \right\|^2 .$$

We report

$$S_{\text{cent}}(C) = \frac{\min_{M \neq C} D_{\text{cent}}(M, C)}{W(C)}, \qquad S_{\text{pair}}(C) = \frac{\min_{M \neq C} D_{\text{pair}}(M, C)}{W(C)}.$$

Values above one indicate that the nearest competing basin is farther away than the model's own endpoint spread.

Silhouette analysis. We compute silhouette scores on self-play endpoints twice: once using model identity as the label and once using topic identity as the label. A higher silhouette under model labels than under topic labels indicates that endpoints are organized more strongly by model identity than by initial topic. Significance is assessed with 1000 random label permutations.

Model	Within-var.	F-ratio	Nearest centroid model	S_{cent}	Nearest pair model	S_{pair}
Grok 4.1	0.0035	23.83	Nemotron	9.44	Nemotron	10.44
Gemini Flash	0.0037	22.60	Qwen 3.5	1.70	Qwen 3.5	2.70
GPT-4o mini	0.0040	20.88	GPT-4.1 nano	7.15	GPT-4.1 nano	8.15
Nemotron	0.0042	19.76	Gemini Flash	5.50	Gemini Flash	6.50
Qwen 3.5	0.0048	17.41	Gemini Flash	1.08	Gemini Flash	2.08
Claude Haiku	0.0070	11.76	Claude Opus	4.32	Claude Opus	5.32
GPT-4.1 nano	0.0092	9.04	GPT-4o mini	2.53	GPT-4o mini	3.53
Claude Opus	0.0156	5.31	Nemotron	1.19	Nemotron	2.19
Between-model var.						0.0827

Table 3. Nearest-rival endpoint variance decomposition in the combined 2-D PCA space, pooling self-play and mixed-play endpoints and averaging over roles. Even under nearest-rival criteria, every model remains locally separated from its closest competing basin.

Model	$\sigma^2(t{=}1)$	$\sigma^2(t{=}T)$	CR
GPT-4.1 nano	0.0027	0.0096	3.515
Grok 4.1	0.0027	0.0140	5.103
Qwen 3.5	0.0027	0.0171	6.217
GPT-4o mini	0.0027	0.0176	6.420
Claude Opus	0.0032	0.0212	6.563
Gemini Flash	0.0027	0.0183	6.659
Nemotron	0.0027	0.0212	7.715
Claude Haiku	0.0027	0.0221	8.041

Table 4. Within-model topic spread at turn 1 and the final turn in the 2-D self-play PCA space. The contraction ratio (CR) is $\sigma^2(t{=}T)/\sigma^2(t{=}1)$. All values exceed 1, indicating divergence across topics rather than contraction toward a point attractor.

Mixed-play quantities. The main text defines the topic-matched mixed-play quantities: partnerward pull α in Eq. 3, off-axis drift $\delta^{\perp}$ in Eq. 4, null-corrected excess in Eq. 5, pair contraction C in Eq. 6, and dominance D in Eq. 7. Appendix tables report topic averages for these quantities. Low α is rigidity-like behavior, high cross-directional α corresponds to influence on the partner, positive C indicates pairwise consensus-like contraction, and positive $D_{A,B}$ means that model A exerts stronger directional pull than model B in that pair.

C.3. Self-Play Basin Diagnostics

Tables 3 and 4 report the full nearest-rival and endpoint-spread diagnostics summarized in the main text. These results support the claim that self-play settles into broad, model-conditioned endpoint basins rather than a single global point attractor.

C.4. Endpoint Hull and Silhouette Checks

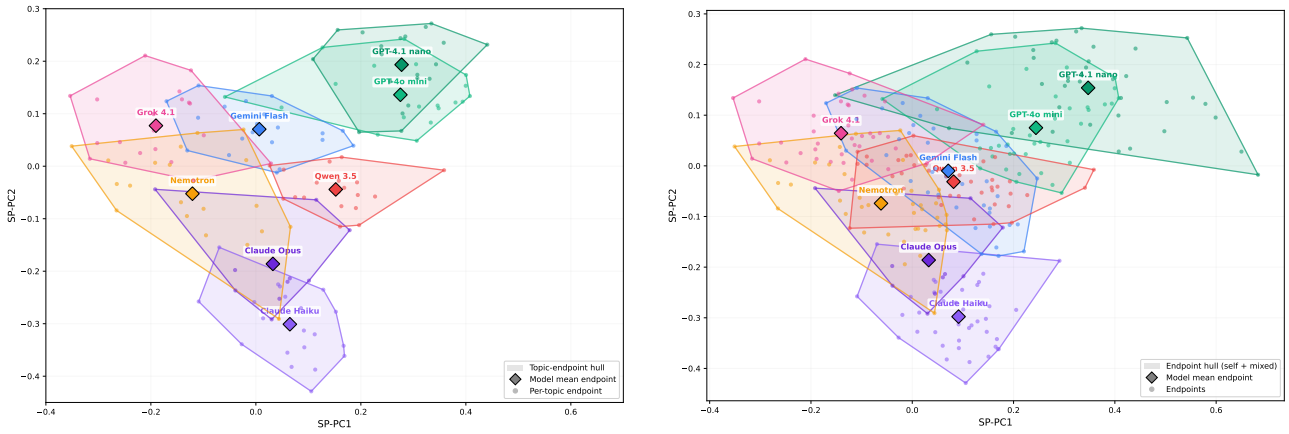
Figure 10 compares self-play-only endpoint hulls with hulls after pooling self-play and mixed-play runs in the same 2-D self-play PCA space. Adding mixed-play increases overlap, as expected, but the clusters do not collapse into a single undifferentiated cloud. Table 5 gives a complementary silhouette analysis: model-identity clustering is weak but positive, whereas topic clustering is negative.

C.5. Mixed-Play Model-Level Decomposition

Table 6 gives the full model-level decomposition used in the main mixed-play analysis. The table consolidates the partnerward-pull, off-axis, null, and excess quantities to avoid treating raw off-axis drift as interaction-specific without the topic-spread correction.

C.6. Centroid-Level Mixed-Play Visualization

The main text reports topic-matched pair contraction and dominance in Table 2. Figure 11 retains the earlier centroid-level visualization as supplementary intuition only.



(a) Self-play only

(b) Self-play and mixed-play pooled

Figure 10. Comparison of endpoint convex hulls in the 2-D self-play PCA space. Adding mixed-play runs increases overlap between some model-conditioned regions, but the overall structure still shows partial separation rather than convergence to one shared basin.

Label source	Silhouette	Permutation p
Model identity	0.0659	< 0.001
Topic	-0.0324	< 0.001

Table 5. Silhouette analysis of endpoint clustering in the 2-D self-play PCA space. Positive silhouette for model identity and negative silhouette for topic indicate that endpoints are organized by model rather than by topic.

C.7. Self-Play Topic-Spread Null for Off-Axis Drift

The null baseline preserves the same same-topic pair axis $(s_{A,k}, s_{B,k})$ but replaces the observed mixed endpoint $m_{A|B,k}$ with a same-model, same-role self-play endpoint from another topic $k' \neq k$. The main text summarizes the observed–null comparison; this appendix records the supplementary diagnostic table and figures.

C.8. Legacy Rigidity/Influence Robustness

The main text uses the topic-matched pair-axis decomposition as the primary analysis. The older rigidity and influence diagnostics are retained here only as coarse robustness checks, because they conflate along-axis movement, off-axis movement, topic spread, and centroid aggregation.

D. Details About Behavioral Signals

To complement the geometric analysis, we characterize each conversation using a set of message-level behavioral signals. These signals are summarized in Table 10. Most are obtained from task-specific LLM judges, which output either ordinal labels, continuous scores, or normalized category weights. In addition, we use a separate emotion classifier to estimate fine-grained affective content. Together, these signals provide a behavioral interpretation of the endpoint basins identified in embedding space: the geometry describes where conversations settle, while the behavioral signals describe the discourse regimes associated with those regions.

For each scalar signal, we compute two model-level summaries. The first is the model's *mean signature*, which measures its average tendency to exhibit the behavior across messages. The second is its *temporal signature*, defined as the late-turn mean minus the early-turn mean, using turns 18–20 and 1–3 respectively. This captures whether a model becomes more or less likely to exhibit a behavior over the course of a conversation. For categorical signals such as argument type and speech

Table 6. Model-level topic-matched mixed-play endpoint decomposition in the full topic-centered 384-D embedding space. Partnerward pull α measures movement along the line between same-topic self-play endpoints. Observed $\delta^{\perp}$ is raw off-axis drift; Null is the self-play topic-spread baseline; Excess is observed minus null; Ratio is observed/null; Interp. is the fraction of endpoints with $0 < \alpha < 1$.

Model	α	Obs. $\delta^{\perp}$	Null	Excess	Ratio	Interp.
Gemini Flash	0.540	0.869	0.679	0.190	1.32	98%
GPT-4o mini	0.442	0.773	0.650	0.123	1.24	99%
Qwen 3.5	0.357	0.773	0.667	0.106	1.17	99%
Grok 4.1	0.517	0.888	0.876	0.013	1.01	100%
Nemotron	0.588	0.871	0.866	0.005	1.01	100%
Claude Haiku	0.266	0.734	0.748	-0.014	0.99	99%
GPT-4.1 nano	0.665	0.788	0.950	-0.162	0.84	100%

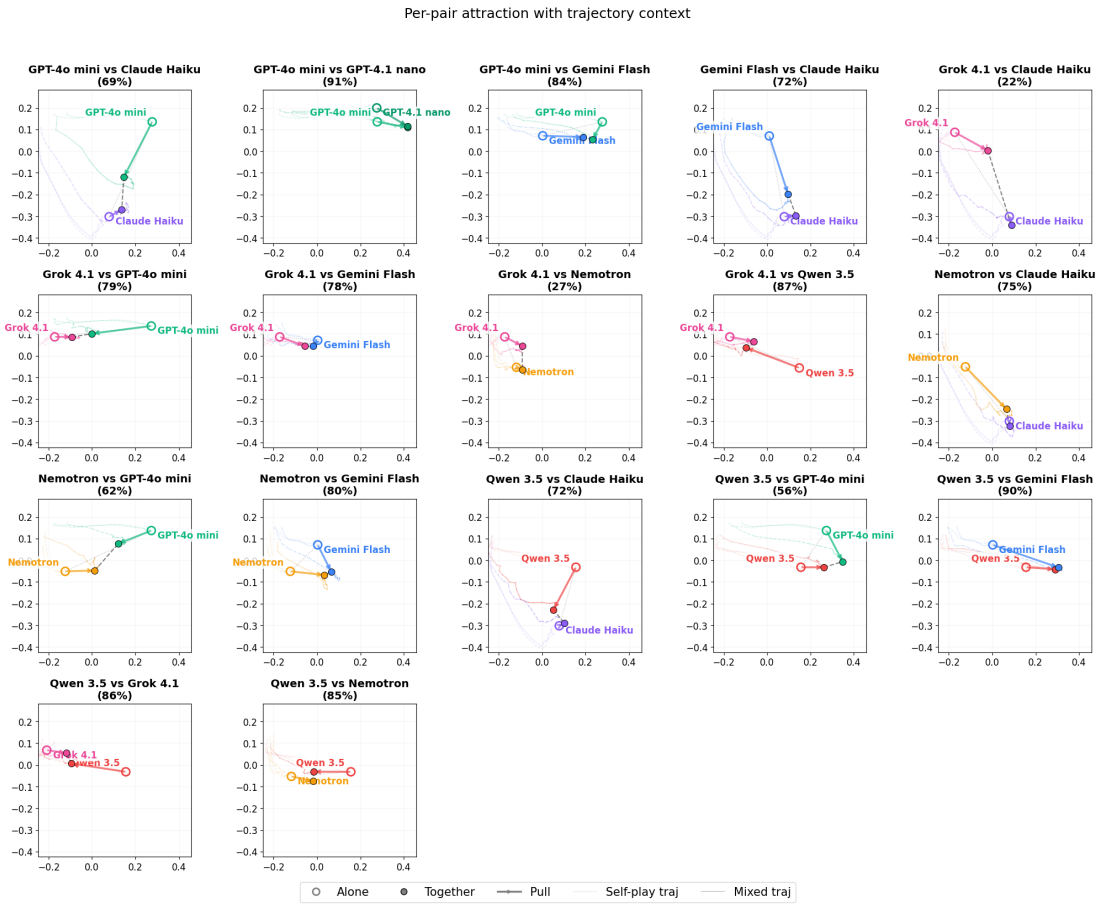


Figure 11. Centroid-level visualization of pairwise mixed-play displacement. Each panel compares a model pair's mixed-play endpoint centroids with the two corresponding self-play centroid regions.

act, we apply the same aggregation procedure to each category proportion.

We also estimate each model's influence on its partners. For each influencing model, we compare an affected partner's behavior in mixed-play against that same partner's behavior in self-play. The *mean influence* measures how much the influencing model raises or lowers the partner's average behavioral signal. The *delta influence* measures whether the influencing model causes the partner's behavior to increase more, or less, over time relative to self-play. These analyses separate a model's own behavioral signature from the behavioral shifts it induces in other models.

Space	Obs. $\delta^{\perp}$	Null $\delta^{\perp}$	Excess	Ratio	p
384-D topic-centered	0.817	0.752	0.065	1.09	0.0005

Table 7. Global comparison between observed mixed-play off-axis drift and the self-play topic-spread null. This diagnostic is used for the null-corrected off-axis analysis in the main text.

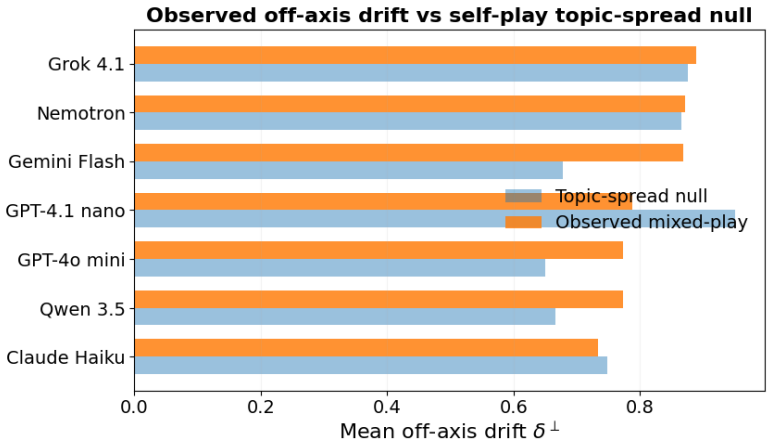


Figure 12. Observed and null off-axis drift by model.

D.1. Temporal Drift from Debate to Affiliation

The main text focuses on model-specific behavioral signatures and asymmetric trait transfer. Here we report a complementary global temporal trend: across conversations, discourse shifts away from explicit contestation and toward affiliation.

Figure 14 shows a broad movement away from explicit contestation. From early turns 0–2 to late turns 18–20, agreement rises from 0.279 to 0.688, rebuttal falls from 0.157 to 0.061, elaboration rises from 0.156 to 0.354, and positivity rises from 0.156 to 0.479. In parallel, negativity drops from 0.311 to 0.128, hedging drops from 0.387 to 0.098, and the rationality-coded score shifts from 0.411 to -0.188. Thus, late-stage conversations are generally more agreeable, more positive, and less explicitly adversarial. We treat this as a broad temporal background pattern; the main behavioral evidence concerns which model-specific traits define self-play basins and which traits transfer asymmetrically in mixed-play.

D.2. Judge-Based Discourse Traits

The LLM-judge traits cover agreement, rationality, sentiment, flattery, hedging, force, assertiveness, argument type, and speech act. Scalar traits are mapped to numerical scores before aggregation. Argument type and speech act instead produce distributions over categories, which we aggregate as category proportions. Table 10 gives the prompt-level definition and downstream scale for each judged signal, and Table 11 gives the category definitions for argument type and speech act.

D.3. Classifier-Based Emotion Traits

We evaluate emotion at the message level using sentence-weighted emotion distribution scoring with the HuggingFace model `SamLowe/roberta-base-go_emotions`. Rather than assigning one emotion label to the full message in a single pass, we first score each sentence independently. We then aggregate the sentence-level emotion distributions into a message-level distribution using a character-length weighted average:

$$p(e \mid x) = \sum_{s \in x} \frac{|s|}{\sum_{s' \in x} |s'|} p(e \mid s), \tag{8}$$

where x is the full message, s indexes its sentences, $|s|$ is the character length of sentence s, and $p(e \mid s)$ is the classifier's predicted probability for emotion e on that sentence. This procedure makes the estimate more robust for longer messages that contain multiple local emotional cues.

The 28 GoEmotions labels are: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire,

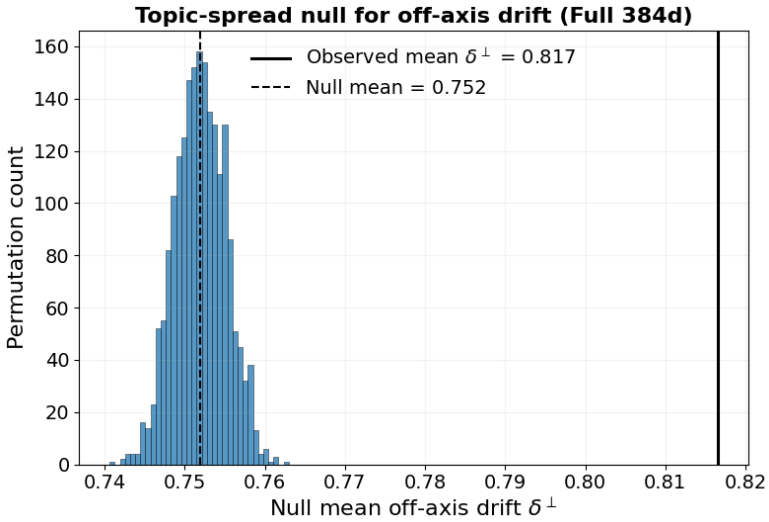


Figure 13. Global self-play topic-spread null distribution for off-axis drift. The vertical marker indicates the observed mean.

Model	Rigidity	Influence
Claude Haiku	0.5386	0.3082
Grok 4.1	0.3564	0.1792
Qwen 3.5	0.3298	0.1801
Nemotron	0.3028	0.1451
GPT-4o mini	0.2600	0.1580
Gemini Flash	0.0636	-0.0897
GPT-4.1 nano	0.0253	-0.0119

Table 8. Supplementary Euclidean rigidity and influence in the original 384-D embedding space. Positive influence indicates pull toward a model's self-play endpoint; negative influence indicates movement away from it. These values are retained as coarse model-level summaries, while the main text uses the pair-axis decomposition to distinguish along-axis pull from off-axis drift.

disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, and neutral.

D.4. Influence on Partners Heatmaps

Figure 15 provides the most direct view of influence on partners across all influencing-model/affected-partner combinations. The heatmaps are oriented with influencing model on the y-axis and affected partner model on the x-axis, so each cell answers: *how much does influencing model M change affected partner model S, relative to S in self-play?* Formally, the mean-influence cell is

$$\text{cell}(M, S) = \mathbb{E}[S \text{ paired with } M] - \mathbb{E}[S \text{ in self-play}],$$

with the diagonal omitted because self-play is the baseline condition. Read row-wise, each row therefore summarizes what a given influencing model does to its affected partners. This is exactly the quantity summarized by the accompanying bar plots: the bar for influencing model M is simply the average of that row over all $S \neq M$.

The same logic applies to temporal change. For each influencing-model/affected-partner pair we compute a delta influence,

$$\text{cell}_{\Delta}(M, S) = \big[\text{late}(S \text{ paired with } M) - \text{early}(S \text{ paired with } M)\big] - \big[\text{late}(S \text{ self-play}) - \text{early}(S \text{ self-play})\big],$$

so positive values indicate that influencing model M makes affected partner S increase more over time than S would in self-play, and negative values indicate a damped or reversed time trend. The corresponding delta-influence bars are again row averages over the off-diagonal cells. Throughout this analysis we keep the current non-role-specific aggregation, pooling across available role rows after the standard dataset filters, so the heatmaps reflect overall influence on partners rather than role-conditioned ones.

Table 9. Full comparison of rigidity and influence across Euclidean and projected formulations, evaluated in both the original 384-D embedding space and the 2-D self-play PCA space. Positive influence indicates pull toward a model's self-play endpoint; negative influence indicates movement away from it.

	Rigidity				Influence			
Model	Euc 384d	Euc 2d	Proj 384d	Proj 2d	Euc 384d	Euc 2d	Proj 384d	Proj 2d
Claude Haiku	**0.5386**	**0.8940**	**0.8582**	**0.9951**	**0.3082**	**0.6326**	**0.6332**	**0.6547**
Grok 4.1	0.3564	0.5958	0.7475	0.6342	0.1792	0.3138	0.3839	0.3199
Qwen 3.5	0.3298	0.3199	0.6218	0.5874	0.1801	0.2015	0.4790	0.5038
Nemotron	0.3028	0.4767	0.6405	0.6159	0.1451	0.2371	0.3652	0.3262
GPT-4o mini	0.2600	0.1792	0.5353	0.7192	0.1580	0.0230	0.3362	0.4834
Gemini Flash	0.0636	0.1745	0.3461	0.3208	-0.0897	0.0625	0.1480	0.1665
GPT-4.1 nano	0.0253	-1.5711	0.2356	-0.3967	-0.0119	-1.6310	0.1870	-0.4863
Mean	0.2681	0.1527	0.5693	0.4966	0.1241	-0.0229	0.3618	0.2812

This orientation is useful because the baseline belongs to the affected partner, not the influencing model. For that reason, the self-play calibration values are best interpreted as affected-partner-specific column annotations rather than as entries in the influence grid itself. Conceptually, each affected partner column can be paired with a small companion row giving its self-play baseline, after which the heatmap cells show how each influencing model pushes that affected partner above or below its own baseline. In this layout, strong positive rows identify influencing models that consistently induce a trait across others, whereas mixed-sign rows indicate more selective interaction effects that depend on which affected partner is being perturbed.

For example, the meta-commentary row for Claude Haiku is strongly positive across most other speakers, matching the main-text claim that Claude induces a meta-commentary-oriented discourse regime. By contrast, the flattery-inducing rows are strongest for Gemini Flash Lite, GPT-4o mini, and Qwen 3.5, showing that these models tend to push their partners toward more socially appreciative late-stage behavior rather than toward reflective process commentary.

Table 12. General scalar metrics. Cells with largest absolute values are in boldfaces

measure	model	mean	partner_mean_shift	early	late	delta	partner_delta_shift
assertiveness	Claude Haiku	0.346	-0.118	0.445	0.286	-0.159	**-0.274**
	GPT-4o mini	0.278	0.007	0.384	0.235	-0.149	0.041
	Gemini Flash Lite	0.297	-0.012	0.383	0.251	-0.132	-0.046
	Grok 4.1	0.661	0.111	0.704	0.539	-0.165	0.160
	Nemotron	**0.745**	**0.244**	0.648	0.752	0.104	0.265
	Qwen 3.5	0.518	0.056	0.635	0.420	**-0.215**	-0.012
force	Claude Haiku	0.582	-0.102	0.612	0.627	0.015	**-0.290**
	GPT-4o mini	0.483	0.009	0.566	0.425	**-0.141**	-0.018
	Gemini Flash Lite	0.568	-0.001	0.602	0.530	-0.072	-0.081
	Grok 4.1	0.833	0.110	0.836	0.795	-0.041	0.179
	Nemotron	**0.868**	**0.174**	0.798	0.864	0.065	0.158
	Qwen 3.5	0.733	0.025	0.753	0.697	-0.056	-0.040
flattery	Claude Haiku	0.281	-0.040	0.350	0.165	-0.184	**-0.201**
	GPT-4o mini	0.516	0.080	0.354	0.532	0.179	0.110
	Gemini Flash Lite	**0.518**	**0.116**	0.336	0.540	0.204	0.096
	Grok 4.1	0.252	-0.041	0.352	0.145	**-0.207**	-0.069
	Nemotron	0.345	-0.020	0.404	0.281	-0.123	-0.109
	Qwen 3.5	0.484	0.097	0.352	0.535	0.182	0.117
negativity	Claude Haiku	0.289	0.067	0.371	0.241	-0.130	0.076
	GPT-4o mini	0.069	-0.015	0.178	0.039	-0.139	0.005

Continued on next page

measure	model	mean	partner_mean_shift	early	late	delta	partner_delta_shift
	Gemini Flash Lite	0.176	-0.006	0.306	0.119	-0.188	0.029
	Grok 4.1	**0.305**	**0.070**	0.413	0.317	-0.096	**0.155**
	Nemotron	0.144	0.033	0.219	0.117	-0.102	0.117
	Qwen 3.5	0.200	0.041	0.328	0.121	**-0.206**	-0.065
neutrality	Claude Haiku	0.503	0.108	0.503	0.474	-0.029	0.093
	GPT-4o mini	0.299	-0.116	0.460	0.256	**-0.204**	-0.053
	Gemini Flash Lite	0.338	-0.039	0.442	0.269	-0.173	-0.027
	Grok 4.1	0.525	0.137	0.471	0.494	0.023	0.188
	Nemotron	**0.680**	**0.159**	0.588	0.718	0.131	**0.281**
	Qwen 3.5	0.430	-0.092	0.515	0.376	-0.139	-0.051
positivity	Claude Haiku	0.208	-0.176	0.126	0.285	0.159	-0.169
	GPT-4o mini	**0.633**	0.131	0.363	0.705	0.343	0.048
	Gemini Flash Lite	0.485	0.045	0.252	0.612	**0.361**	-0.002
	Grok 4.1	0.170	**-0.206**	0.116	0.189	0.074	-0.343
	Nemotron	0.176	-0.192	0.193	0.165	-0.028	**-0.398**
	Qwen 3.5	0.370	0.051	0.157	0.503	0.345	0.116
rationality	Claude Haiku	0.182	0.009	0.458	0.010	-0.449	-0.039
	GPT-4o mini	-0.105	-0.191	0.406	-0.324	-0.730	-0.208
	Gemini Flash Lite	0.062	-0.194	0.375	-0.099	-0.475	-0.250
	Grok 4.1	**0.257**	0.123	0.502	0.066	-0.437	0.025
	Nemotron	-0.118	**-0.223**	0.369	-0.328	-0.697	-0.190
	Qwen 3.5	-0.103	-0.215	0.352	-0.413	**-0.765**	**-0.419**
agreement	Claude Haiku	0.363	-0.076	0.115	0.533	0.419	0.131
	GPT-4o mini	0.812	0.060	0.490	0.848	0.358	**0.295**
	Gemini Flash Lite	0.696	0.039	0.373	0.798	0.425	0.278
	Grok 4.1	0.194	**-0.278**	-0.123	0.358	0.481	-0.063
	Nemotron	**0.877**	0.167	0.796	0.904	0.108	-0.062
	Qwen 3.5	0.622	-0.109	0.146	0.842	**0.696**	0.231

Table 13. Argument-type labels. Cells with largest absolute values are in boldfaces

measure	model	mean	partner_mean_shift	early	late	delta	partner_delta_shift
analogy	Claude Haiku	0.005	-0.002	0.013	0.001	-0.012	**-0.014**
	GPT-4o mini	0.001	-0.001	0.001	0.001	0.000	-0.003
	Gemini Flash Lite	0.006	0.005	0.008	0.002	-0.007	0.004
	Grok 4.1	0.023	0.005	0.034	0.009	**-0.025**	0.004
	Nemotron	0.015	0.001	0.017	0.011	-0.006	0.000
	Qwen 3.5	**0.027**	**0.009**	0.034	0.018	-0.016	0.006
concession	Claude Haiku	**0.130**	**0.098**	0.115	0.126	0.011	0.018
	GPT-4o mini	0.106	-0.041	0.189	0.044	**-0.145**	-0.047
	Gemini Flash Lite	0.115	-0.011	0.148	0.072	-0.076	-0.034
	Grok 4.1	0.047	0.004	0.071	0.022	-0.049	-0.029
	Nemotron	0.033	-0.046	0.079	0.011	-0.068	**-0.060**
	Qwen 3.5	0.088	0.009	0.084	0.074	-0.011	-0.031
counter_evidence	Claude Haiku	0.047	-0.021	0.189	0.003	**-0.186**	-0.015
	GPT-4o mini	0.025	-0.019	0.116	0.005	-0.111	-0.057
	Gemini Flash Lite	0.039	-0.015	0.143	0.013	-0.130	-0.043
	Grok 4.1	**0.137**	**0.037**	0.242	0.084	-0.158	-0.049
	Nemotron	0.033	-0.023	0.108	0.009	-0.099	-0.007

Continued on next page

measure	model	mean	partner_mean_shift	early	late	delta	partner_delta_shift
	Qwen 3.5	0.057	0.017	0.187	0.008	-0.179	**-0.066**
elaboration	Claude Haiku	0.104	**-0.228**	0.125	0.054	-0.071	-0.129
	GPT-4o mini	0.422	0.113	0.307	0.376	0.069	0.263
	Gemini Flash Lite	0.344	0.010	0.259	0.337	0.077	0.193
	Grok 4.1	0.340	-0.011	0.177	0.438	**0.260**	**0.268**
	Nemotron	**0.482**	0.170	0.458	0.442	-0.017	0.137
	Qwen 3.5	0.307	-0.045	0.173	0.402	0.229	0.222
meta_commentary	Claude Haiku	**0.345**	**0.151**	0.065	0.476	**0.411**	**0.161**
	GPT-4o mini	0.055	-0.002	0.012	0.028	0.016	0.021
	Gemini Flash Lite	0.082	0.021	0.016	0.048	0.032	0.025
	Grok 4.1	0.074	0.025	0.022	0.073	0.051	0.076
	Nemotron	0.118	0.035	0.035	0.120	0.085	0.038
	Qwen 3.5	0.124	0.003	0.021	0.149	0.128	0.050
phatic_bridge	Claude Haiku	0.136	-0.005	0.024	0.270	0.246	0.108
	GPT-4o mini	**0.285**	-0.013	0.015	0.506	**0.491**	-0.053
	Gemini Flash Lite	0.228	-0.013	0.025	0.430	0.404	-0.050
	Grok 4.1	0.072	**-0.146**	0.033	0.108	0.075	**-0.273**
	Nemotron	0.166	-0.091	0.016	0.314	0.298	-0.162
	Qwen 3.5	0.146	-0.049	0.013	0.228	0.215	-0.105
rebuttal	Claude Haiku	0.112	-0.017	0.271	0.033	-0.238	-0.079
	GPT-4o mini	0.054	-0.030	0.184	0.027	-0.157	-0.103
	Gemini Flash Lite	0.120	-0.014	0.259	0.070	-0.189	-0.091
	Grok 4.1	**0.179**	0.047	0.255	0.156	-0.098	-0.053
	Nemotron	0.031	**-0.060**	0.073	0.017	-0.055	-0.005
	Qwen 3.5	0.107	0.025	0.272	0.030	**-0.242**	**-0.109**
reframing	Claude Haiku	0.117	0.022	0.189	0.036	-0.153	-0.047
	GPT-4o mini	0.051	-0.006	0.173	0.013	**-0.160**	-0.020
	Gemini Flash Lite	0.065	0.016	0.138	0.028	-0.109	-0.005
	Grok 4.1	0.124	**0.037**	0.159	0.109	-0.050	0.056
	Nemotron	0.121	0.014	0.212	0.076	-0.136	**0.059**
	Qwen 3.5	**0.140**	0.029	0.209	0.091	-0.118	0.035

Table 14. Speech-act labels. Cells with largest absolute values are in boldfaces

measure	model	mean	partner_mean_shift	early	late	delta	partner_delta_shift
assertive	Claude Haiku	0.615	-0.093	0.795	0.392	**-0.403**	**-0.240**
	GPT-4o mini	0.586	**-0.099**	0.854	0.463	-0.391	-0.122
	Gemini Flash Lite	0.697	-0.060	0.844	0.615	-0.229	-0.067
	Grok 4.1	**0.749**	-0.028	0.737	0.764	0.027	0.024
	Nemotron	0.690	-0.033	0.770	0.653	-0.117	0.018
	Qwen 3.5	0.620	-0.066	0.827	0.490	-0.338	-0.126
commissive	Claude Haiku	**0.095**	0.007	0.019	0.223	**0.204**	-0.000
	GPT-4o mini	0.075	0.002	0.017	0.088	0.071	0.004
	Gemini Flash Lite	0.028	-0.010	0.008	0.034	0.025	-0.019
	Grok 4.1	0.020	0.009	0.029	0.013	-0.016	-0.018
	Nemotron	0.049	0.007	0.040	0.050	0.010	0.010
	Qwen 3.5	0.047	**-0.015**	0.019	0.051	0.032	**-0.030**
declaration	Claude Haiku	0.020	0.010	0.000	0.060	**0.060**	**0.040**
	GPT-4o mini	0.001	-0.002	0.000	0.004	0.004	-0.001

Continued on next page

measure	model	mean	partner_mean_shift	early	late	delta	partner_delta_shift
	Gemini Flash Lite	0.003	-0.003	0.000	0.008	0.008	-0.001
	Grok 4.1	0.010	0.001	0.001	0.017	0.017	0.003
	Nemotron	**0.029**	**0.014**	0.003	0.051	0.048	0.023
	Qwen 3.5	0.012	-0.001	0.000	0.031	0.030	-0.003
directive	Claude Haiku	0.081	-0.006	0.079	0.072	-0.007	0.017
	GPT-4o mini	0.054	0.040	0.041	0.072	0.031	0.050
	Gemini Flash Lite	0.022	0.009	0.024	0.026	0.002	0.007
	Grok 4.1	0.108	0.037	0.115	0.091	-0.024	0.041
	Nemotron	**0.139**	**0.068**	0.093	0.148	**0.056**	**0.083**
	Qwen 3.5	0.082	0.012	0.054	0.099	0.045	0.052
expressive	Claude Haiku	0.185	**0.074**	0.107	0.242	0.135	**0.166**
	GPT-4o mini	**0.284**	0.060	0.087	0.372	**0.284**	0.073
	Gemini Flash Lite	0.250	0.064	0.124	0.317	0.193	0.079
	Grok 4.1	0.112	-0.019	0.118	0.114	-0.004	-0.047
	Nemotron	0.087	-0.058	0.094	0.087	-0.007	-0.135
	Qwen 3.5	0.239	0.071	0.099	0.329	0.230	0.109

Table 15. Emotion-analysis labels.

measure	model	mean	partner_mean_shift	early	late	delta	partner_delta_shift
admiration	Claude Haiku	0.072	-0.039	0.046	0.097	0.051	-0.092
	GPT-4o mini	0.157	0.049	0.064	0.161	0.097	0.033
	Gemini Flash Lite	**0.162**	0.036	0.059	0.213	**0.153**	0.044
	Grok 4.1	0.031	-0.061	0.037	0.029	-0.008	-0.083
	Nemotron	0.022	**-0.065**	0.028	0.019	-0.008	**-0.107**
	Qwen 3.5	0.094	0.032	0.026	0.136	0.110	0.039
amusement	Claude Haiku	0.001	**0.000**	0.001	0.002	0.000	**0.001**
	GPT-4o mini	0.001	0.000	0.001	0.002	**0.001**	0.000
	Gemini Flash Lite	0.001	0.000	0.001	0.001	0.000	0.000
	Grok 4.1	**0.002**	-0.000	0.002	0.001	-0.000	-0.000
	Nemotron	0.002	0.000	0.001	0.002	0.000	-0.000
	Qwen 3.5	0.002	0.000	0.001	0.002	0.000	0.000
anger	Claude Haiku	**0.004**	**0.001**	0.003	0.006	**0.003**	0.001
	GPT-4o mini	0.001	0.001	0.001	0.002	0.000	**0.002**
	Gemini Flash Lite	0.002	0.000	0.002	0.002	-0.001	0.001
	Grok 4.1	0.003	0.001	0.003	0.004	0.001	0.001
	Nemotron	0.003	0.001	0.002	0.003	0.000	0.001
	Qwen 3.5	0.003	0.001	0.003	0.002	-0.000	0.000
annoyance	Claude Haiku	**0.026**	**0.005**	0.029	0.028	-0.002	0.002
	GPT-4o mini	0.008	0.002	0.012	0.006	-0.006	0.003
	Gemini Flash Lite	0.013	0.000	0.020	0.010	-0.010	0.002
	Grok 4.1	0.020	0.004	0.032	0.019	**-0.013**	**0.008**
	Nemotron	0.013	0.004	0.018	0.011	-0.007	0.005
	Qwen 3.5	0.017	0.004	0.024	0.012	-0.012	-0.004
approval	Claude Haiku	0.180	-0.105	0.185	0.147	-0.037	**-0.175**
	GPT-4o mini	**0.360**	-0.006	0.415	0.306	**-0.109**	-0.069
	Gemini Flash Lite	0.355	-0.042	0.331	0.345	0.014	-0.085
	Grok 4.1	0.080	-0.061	0.125	0.076	-0.049	-0.032
	Nemotron	0.103	**-0.111**	0.171	0.072	-0.099	-0.148

Continued on next page

measure	model	mean	partner_mean_shift	early	late	delta	partner_delta_shift
	Qwen 3.5	0.204	-0.068	0.206	0.193	-0.012	-0.092
caring	Claude Haiku	0.034	0.002	0.017	0.043	0.026	0.016
	GPT-4o mini	**0.056**	0.010	0.037	0.061	0.024	-0.006
	Gemini Flash Lite	0.037	0.008	0.029	0.052	0.023	0.009
	Grok 4.1	0.014	**-0.045**	0.006	0.020	0.013	**-0.072**
	Nemotron	0.006	-0.031	0.007	0.005	-0.002	-0.051
	Qwen 3.5	0.042	0.019	0.011	0.078	**0.067**	0.036
confusion	Claude Haiku	**0.042**	0.005	0.031	0.028	-0.003	**0.013**
	GPT-4o mini	0.006	-0.008	0.013	0.004	-0.010	-0.008
	Gemini Flash Lite	0.011	**-0.010**	0.022	0.006	**-0.017**	0.003
	Grok 4.1	0.026	0.003	0.027	0.014	-0.012	0.012
	Nemotron	0.007	-0.002	0.010	0.006	-0.004	0.012
	Qwen 3.5	0.010	-0.006	0.015	0.006	-0.009	-0.008
curiosity	Claude Haiku	0.028	0.002	0.034	0.012	**-0.022**	0.009
	GPT-4o mini	0.005	0.002	0.005	0.006	0.001	0.004
	Gemini Flash Lite	0.006	0.005	0.008	0.004	-0.004	**0.012**
	Grok 4.1	**0.030**	**0.006**	0.036	0.015	-0.021	0.005
	Nemotron	0.007	-0.001	0.009	0.006	-0.003	0.009
	Qwen 3.5	0.009	0.004	0.014	0.007	-0.007	0.003
desire	Claude Haiku	0.014	-0.014	0.007	0.011	0.004	**-0.019**
	GPT-4o mini	**0.026**	0.004	0.016	0.026	0.010	0.002
	Gemini Flash Lite	0.016	0.001	0.011	0.016	0.006	-0.002
	Grok 4.1	0.002	**-0.016**	0.004	0.002	-0.002	-0.017
	Nemotron	0.005	-0.011	0.006	0.005	-0.001	-0.017
	Qwen 3.5	0.018	0.007	0.009	0.021	**0.012**	0.009
disappointment	Claude Haiku	0.018	**0.004**	0.030	0.014	-0.017	0.008
	GPT-4o mini	0.008	0.000	0.021	0.006	-0.015	0.001
	Gemini Flash Lite	0.016	-0.000	0.030	0.010	**-0.020**	0.007
	Grok 4.1	**0.021**	0.002	0.039	0.023	-0.016	**0.014**
	Nemotron	0.008	0.000	0.017	0.006	-0.011	0.001
	Qwen 3.5	0.015	0.002	0.027	0.010	-0.017	-0.005
disapproval	Claude Haiku	**0.066**	0.008	0.080	0.070	-0.009	0.012
	GPT-4o mini	0.013	-0.001	0.031	0.008	-0.023	0.009
	Gemini Flash Lite	0.035	-0.007	0.053	0.026	-0.028	0.002
	Grok 4.1	0.041	**0.014**	0.060	0.045	-0.015	**0.028**
	Nemotron	0.019	-0.001	0.034	0.013	-0.020	0.018
	Qwen 3.5	0.035	0.001	0.058	0.021	**-0.037**	-0.016
disgust	Claude Haiku	0.002	**0.000**	0.002	0.002	-0.000	**0.001**
	GPT-4o mini	0.001	0.000	0.002	0.001	-0.000	-0.000
	Gemini Flash Lite	0.002	0.000	0.002	0.001	**-0.001**	0.000
	Grok 4.1	**0.003**	0.000	0.004	0.003	-0.000	0.001
	Nemotron	0.002	0.000	0.002	0.002	-0.000	0.000
	Qwen 3.5	0.002	0.000	0.002	0.002	-0.001	-0.000
embarrassment	Claude Haiku	**0.001**	**0.000**	0.002	0.001	**-0.001**	**0.000**
	GPT-4o mini	0.001	0.000	0.001	0.001	-0.000	-0.000
	Gemini Flash Lite	0.001	0.000	0.001	0.001	-0.001	0.000
	Grok 4.1	0.001	0.000	0.002	0.001	-0.001	0.000
	Nemotron	0.001	0.000	0.001	0.001	-0.001	0.000
	Qwen 3.5	0.001	0.000	0.001	0.001	-0.000	-0.000
excitement	Claude Haiku	0.003	-0.006	0.002	0.003	0.001	**-0.008**

Continued on next page

measure	model	mean	partner_mean_shift	early	late	delta	partner_delta_shift
	GPT-4o mini	**0.013**	**0.007**	0.004	0.016	**0.012**	0.005
	Gemini Flash Lite	0.010	0.004	0.003	0.014	0.011	0.008
	Grok 4.1	0.003	-0.001	0.001	0.003	0.002	-0.003
	Nemotron	0.003	-0.002	0.002	0.003	0.001	-0.003
	Qwen 3.5	0.007	0.004	0.003	0.009	0.007	0.005
fear	Claude Haiku	0.003	-0.000	0.003	0.003	-0.000	0.003
	GPT-4o mini	0.002	-0.000	0.004	0.001	-0.002	-0.000
	Gemini Flash Lite	**0.004**	**0.001**	0.009	0.002	**-0.006**	0.001
	Grok 4.1	0.003	-0.001	0.004	0.002	-0.002	**0.004**
	Nemotron	0.002	-0.000	0.003	0.002	-0.001	0.001
	Qwen 3.5	0.004	0.001	0.006	0.002	-0.004	-0.001
gratitude	Claude Haiku	0.061	0.043	0.005	0.101	0.096	**0.159**
	GPT-4o mini	**0.117**	0.028	0.013	0.183	**0.170**	0.010
	Gemini Flash Lite	0.060	0.018	0.007	0.107	0.100	0.031
	Grok 4.1	0.012	-0.064	0.010	0.016	0.007	-0.095
	Nemotron	0.008	**-0.070**	0.006	0.005	-0.001	-0.112
	Qwen 3.5	0.052	0.017	0.004	0.082	0.078	0.032
grief	Claude Haiku	**0.001**	**0.000**	0.001	0.003	**0.002**	**0.003**
	GPT-4o mini	0.001	0.000	0.001	0.002	0.001	0.001
	Gemini Flash Lite	0.001	0.000	0.001	0.001	-0.000	0.000
	Grok 4.1	0.001	-0.000	0.001	0.001	-0.000	-0.000
	Nemotron	0.000	-0.000	0.000	0.000	-0.000	-0.000
	Qwen 3.5	0.001	0.000	0.001	0.002	0.001	0.001
joy	Claude Haiku	0.004	-0.014	0.002	0.004	0.002	-0.022
	GPT-4o mini	**0.022**	0.006	0.006	0.030	**0.024**	0.007
	Gemini Flash Lite	0.015	0.007	0.003	0.023	0.019	0.006
	Grok 4.1	0.008	-0.013	0.002	0.008	0.006	-0.028
	Nemotron	0.007	**-0.016**	0.007	0.006	-0.002	**-0.033**
	Qwen 3.5	0.018	0.005	0.004	0.026	0.022	0.015
love	Claude Haiku	0.002	0.000	0.002	0.004	0.003	0.002
	GPT-4o mini	**0.004**	0.001	0.002	0.005	0.003	0.002
	Gemini Flash Lite	0.003	0.001	0.002	0.004	0.002	0.001
	Grok 4.1	0.002	**-0.002**	0.003	0.002	-0.001	**-0.003**
	Nemotron	0.002	0.000	0.002	0.002	-0.001	-0.001
	Qwen 3.5	0.004	0.001	0.002	0.006	**0.004**	0.003
nervousness	Claude Haiku	**0.003**	**-0.001**	0.004	0.003	-0.001	**0.002**
	GPT-4o mini	0.002	0.000	0.005	0.001	-0.004	-0.001
	Gemini Flash Lite	0.003	0.000	0.008	0.002	**-0.007**	0.002
	Grok 4.1	0.001	-0.001	0.003	0.001	-0.002	0.002
	Nemotron	0.001	-0.000	0.001	0.000	-0.001	-0.001
	Qwen 3.5	0.002	0.001	0.003	0.001	-0.002	-0.002
neutral	Claude Haiku	0.497	0.122	0.570	0.457	-0.112	0.065
	GPT-4o mini	0.245	-0.084	0.359	0.226	-0.133	0.018
	Gemini Flash Lite	0.300	-0.016	0.413	0.232	-0.180	-0.020
	Grok 4.1	0.767	0.274	0.657	0.788	0.131	0.314
	Nemotron	**0.818**	**0.308**	0.713	0.864	0.151	**0.421**
	Qwen 3.5	0.487	-0.022	0.605	0.410	**-0.195**	-0.012
optimism	Claude Haiku	0.023	-0.080	0.019	0.021	0.002	-0.116
	GPT-4o mini	**0.119**	0.005	0.094	0.114	0.020	-0.011
	Gemini Flash Lite	0.081	-0.005	0.054	0.091	0.036	-0.024

Continued on next page

measure	model	mean	partner_mean_shift	early	late	delta	partner_delta_shift
	Grok 4.1	0.009	**-0.100**	0.012	0.009	-0.003	**-0.148**
	Nemotron	0.014	-0.083	0.021	0.009	-0.012	-0.118
	Qwen 3.5	0.073	0.004	0.025	0.092	**0.068**	0.019
pride	Claude Haiku	0.002	-0.002	0.001	0.002	0.001	-0.003
	GPT-4o mini	**0.004**	0.001	0.002	0.005	**0.003**	0.000
	Gemini Flash Lite	0.003	0.000	0.002	0.004	0.002	-0.000
	Grok 4.1	0.001	**-0.003**	0.001	0.001	0.000	**-0.004**
	Nemotron	0.001	-0.002	0.001	0.001	-0.000	-0.003
	Qwen 3.5	0.003	0.001	0.001	0.005	0.003	0.002
realization	Claude Haiku	**0.044**	**0.010**	0.043	0.036	-0.007	0.008
	GPT-4o mini	0.034	0.004	0.052	0.025	**-0.028**	0.001
	Gemini Flash Lite	0.039	0.002	0.051	0.033	-0.018	0.004
	Grok 4.1	0.018	-0.003	0.034	0.017	-0.017	0.003
	Nemotron	0.024	0.007	0.034	0.020	-0.014	**0.013**
	Qwen 3.5	0.038	0.001	0.047	0.034	-0.014	0.002
relief	Claude Haiku	0.002	-0.002	0.002	0.002	0.001	-0.003
	GPT-4o mini	**0.005**	0.001	0.004	0.005	0.001	-0.000
	Gemini Flash Lite	0.004	0.000	0.003	0.004	0.001	-0.001
	Grok 4.1	0.001	**-0.003**	0.001	0.001	-0.000	**-0.005**
	Nemotron	0.002	-0.002	0.002	0.001	-0.001	-0.005
	Qwen 3.5	0.004	0.000	0.002	0.005	**0.002**	0.001
remorse	Claude Haiku	**0.002**	**0.001**	0.001	0.002	**0.001**	**0.001**
	GPT-4o mini	0.001	-0.000	0.002	0.002	0.000	-0.001
	Gemini Flash Lite	0.002	0.000	0.002	0.002	-0.000	-0.001
	Grok 4.1	0.001	-0.001	0.002	0.001	-0.001	-0.001
	Nemotron	0.001	-0.000	0.001	0.000	-0.000	-0.001
	Qwen 3.5	0.001	-0.000	0.001	0.002	0.000	-0.000
sadness	Claude Haiku	**0.012**	**0.010**	0.009	0.043	**0.034**	**0.056**
	GPT-4o mini	0.007	0.003	0.009	0.020	0.011	0.012
	Gemini Flash Lite	0.006	0.002	0.013	0.005	-0.008	0.003
	Grok 4.1	0.009	0.001	0.016	0.009	-0.007	0.002
	Nemotron	0.004	0.001	0.006	0.003	-0.003	0.002
	Qwen 3.5	0.009	0.005	0.012	0.018	0.007	0.011
surprise	Claude Haiku	**0.002**	**0.000**	0.003	0.002	**-0.001**	0.000
	GPT-4o mini	0.001	0.000	0.001	0.002	0.001	**-0.002**
	Gemini Flash Lite	0.001	0.000	0.001	0.001	0.000	-0.002
	Grok 4.1	0.001	-0.000	0.001	0.001	-0.000	0.000
	Nemotron	0.001	-0.000	0.001	0.001	-0.000	-0.000
	Qwen 3.5	0.002	0.000	0.001	0.002	0.000	-0.002
dominant_score	Claude Haiku	0.552	0.109	0.570	0.554	-0.016	0.143
	GPT-4o mini	0.456	-0.066	0.455	0.485	0.030	0.003
	Gemini Flash Lite	0.445	-0.014	0.446	0.450	0.004	-0.045
	Grok 4.1	0.774	**0.203**	0.657	0.788	0.131	0.179
	Nemotron	**0.820**	0.177	0.716	0.865	**0.149**	**0.225**
	Qwen 3.5	0.541	-0.041	0.606	0.508	-0.098	-0.028

E. Stance Calibration

To control for the possibility that observed stance trajectories partly reflect models' intrinsic priors rather than interaction effects, we also report calibrated versions of the trajectories. In Figure 16, each model's scores are centered by its corresponding control-condition baseline, so the plotted values reflect deviation from that model's default stance. In Figure 17, scores are centered around the neutral midpoint of the Likert scale (3.0), which provides a common reference

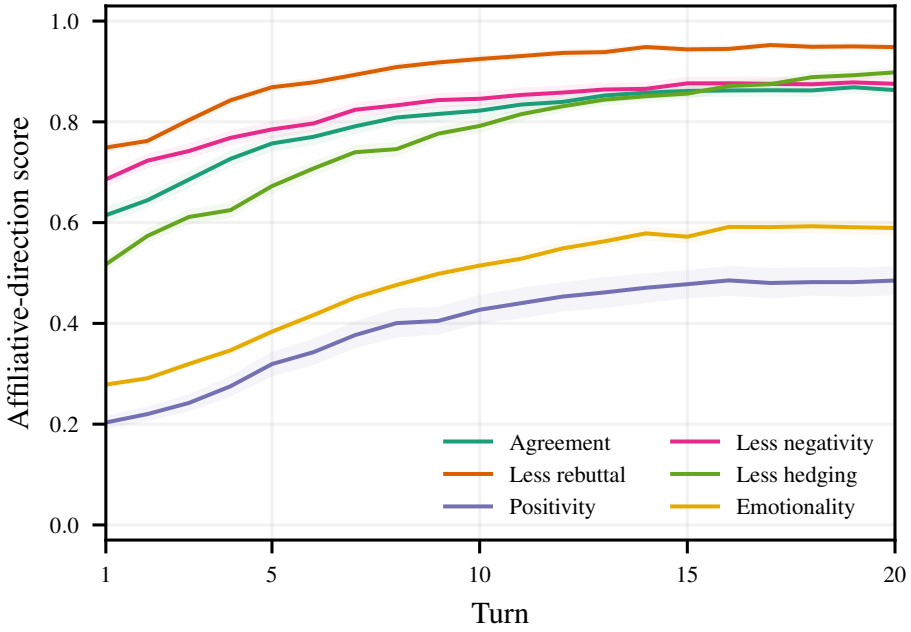


Figure 14. Discourse shifts from debate toward affiliation over the course of conversation. Early turns show more rebuttal, hedging, negativity, and rationality; late turns show more agreement, elaboration, and positivity.

across models. Both views preserve the main qualitative trends, indicating that the mixed-play dynamics are not solely an artifact of fixed model-specific response biases.

F. Additional Lexical and Semantic Dynamics

We summarize the supporting lexical and semantic measurements here. Lexically, we tokenize each response and compute turn-level lexicon entropy from the word-frequency histogram, together with ROUGE-L overlap between consecutive turns as a measure of local lexical repetition. Semantically, we use SBERT embeddings to compute sequential similarity between an agent's consecutive turns, same-turn similarity between the two agents, and topic deviation from the conversation's initialization anchor. These measures serve as supplementary diagnostics of conversational drift and reuse rather than as primary evidence for the main geometric claims.

Stylistic behavior transfer. Feature-level influence plots in Fig. 18 show four examples that stylistic and discourse behaviors can transfer asymmetrically across models. In Fig. 18a–c, Claude Haiku exerts a strong pull on its interlocutors: mixed-play trajectories shift toward Claude-associated patterns in explicit AI-role expression, boldface formatting, and conversation-termination language, even when those behaviors are weak or absent in the affected partner model's self-play. This indicates that Claude does not merely preserve its own style, but can actively reshape the joint interaction along these dimensions. The asymmetry is not universal, however. In Fig. 18d, the appreciativeness feature instead shows Qwen pulling Gemini, demonstrating that feature-specific influence can be dominated by a different model. Together, these cases provide concrete evidence that behavioral transfer in mixed play is directional and feature-dependent rather than evenly shared across participants.

The broader lexical and semantic trends help qualify this stylistic-transfer result. Claude's own semantic similarity decreases in mixed-play relative to its self-play baseline (Fig. 20c), suggesting that while some of its surface-level stylistic markers propagate to interlocutors, Claude itself remains comparatively flexible at the semantic level. This pattern is consistent with an asymmetric interaction in which one model shapes conversational form without fully fixing conversational meaning.

With a sufficient number of turns, we also observe consistent lexical and semantic trends across interactions. Lexicon entropy decreases monotonically with turn index (Fig. 20a), indicating a reduction in vocabulary diversity over time. Token-based overlap, measured by ROUGE-L similarity (Fig. 20b), also generally increases, although the magnitude differs by model family. It rises and stabilizes around 0.67–0.85 for GPT, Gemini, and Nemotron, while it increases more slowly for Grok

Table 10. Definitions of judge-based discourse traits derived from the prompts used in our pipeline. The table summarizes what each judge outputs and how those outputs are converted into the scalar scores or category shares used in aggregation.

Judge output	Definition used in prompts	Downstream scale
Agreement	Degree to which the current turn agrees with the immediately prior turn, from explicit endorsement through partial disagreement to direct rejection, with a `not applicable` option when the prior turn has no clear position.	Five-level ordinal mapped to $[-1, 1]$: $1.0, 0.5, 0, -0.5, -1.0$.
Rationality	Whether the message is framed primarily through logic, evidence, and definitions versus feeling, intuition, or passion.	Five-level ordinal mapped to $[-1, 1]$; higher values are more rational.
Sentiment polarity	Emotional direction of the message, from strongly positive to strongly negative.	Five-level ordinal mapped to $[-1, 1]$.
Sentiment intensity	Strength of emotional expression, from flat or technical language to highly intense affect.	Four-level ordinal mapped to $[0, 1]$: $0, 0.33, 0.66, 1.0$.
Flattery	Praise directed at the interlocutor or the conversation itself rather than substantive topic engagement; the judge also records flattery type and conversational function.	Level mapped to $[0, 1]$; type and function retained categorically.
Hedging	Degree of epistemic qualification, i.e., language that weakens certainty or distances the speaker from full commitment.	Level mapped to $[0, 1]$; count and type retained separately.
Force	How directly and confidently the speaker commits to the claim, from deferential or non-committal to fully staked without hedging.	Continuous $[0, 1]$.
Assertiveness	How strongly the speaker presses the addressee to accept the position, independent of mere claim confidence.	Continuous $[0, 1]$.
Argument-type category weights	Distribution over eight argumentative-move categories relative to the prior turn; detailed category definitions are given in Table 11.	Sparse weights in $[0, 1]$ summing to 1 across categories.
Speech-act category weights	Distribution over five illocutionary-act categories, independent of argumentative role; detailed category definitions are given in Table 11.	Category proportions in $[0, 1]$ summing to 1 across categories.

and Claude, from roughly 0.13 to around 0.2.

At the same time, semantic similarity does not increase in parallel. In several cases it remains flat or even decreases (Fig. 20c), showing that conversations can reuse more lexical material while continuing to express distinct semantic content. Topic deviation also continues to rise over turns (Fig. 20d), indicating that conversations gradually drift away from the initial topic anchor.

F.1. Pairwise Stance Trajectory

Figure 19 shows a simplified visualization of pair-wise stance trajectories, separated per interaction.

Table 11. Label definitions for the categorical outputs of the argument-type and speech-act LLM judges.

Category family	Label	Definition
Argument type	rebuttal	Directly opposes the prior claim with counter-reasoning.
	counter_evidence	Introduces new facts or examples against the prior claim.
	reframing	Accepts the topic but shifts the angle or terms.
	concession	Yields to or incorporates the other position.
	analogy	Uses comparison or metaphor as the primary argumentative tool.
	elaboration	Extends or deepens a prior claim without opposing it.
	meta_commentary	Comments on the conversation structure or process itself.
	phatic_bridge	Social or transitional content with no argumentative load.
Speech act	assertive	Commits the speaker to a proposition being true, e.g., claiming, concluding, or stating.
	directive	Attempts to get the addressee to do something, e.g., requesting, questioning, or challenging.
	commissive	Commits the speaker to a future action, e.g., promising, offering, or planning.
	expressive	Expresses a psychological state, e.g., thanking, apologizing, or welcoming.
	declaration	Changes an institutional state of affairs by being uttered, e.g., ruling, declaring, or firing.

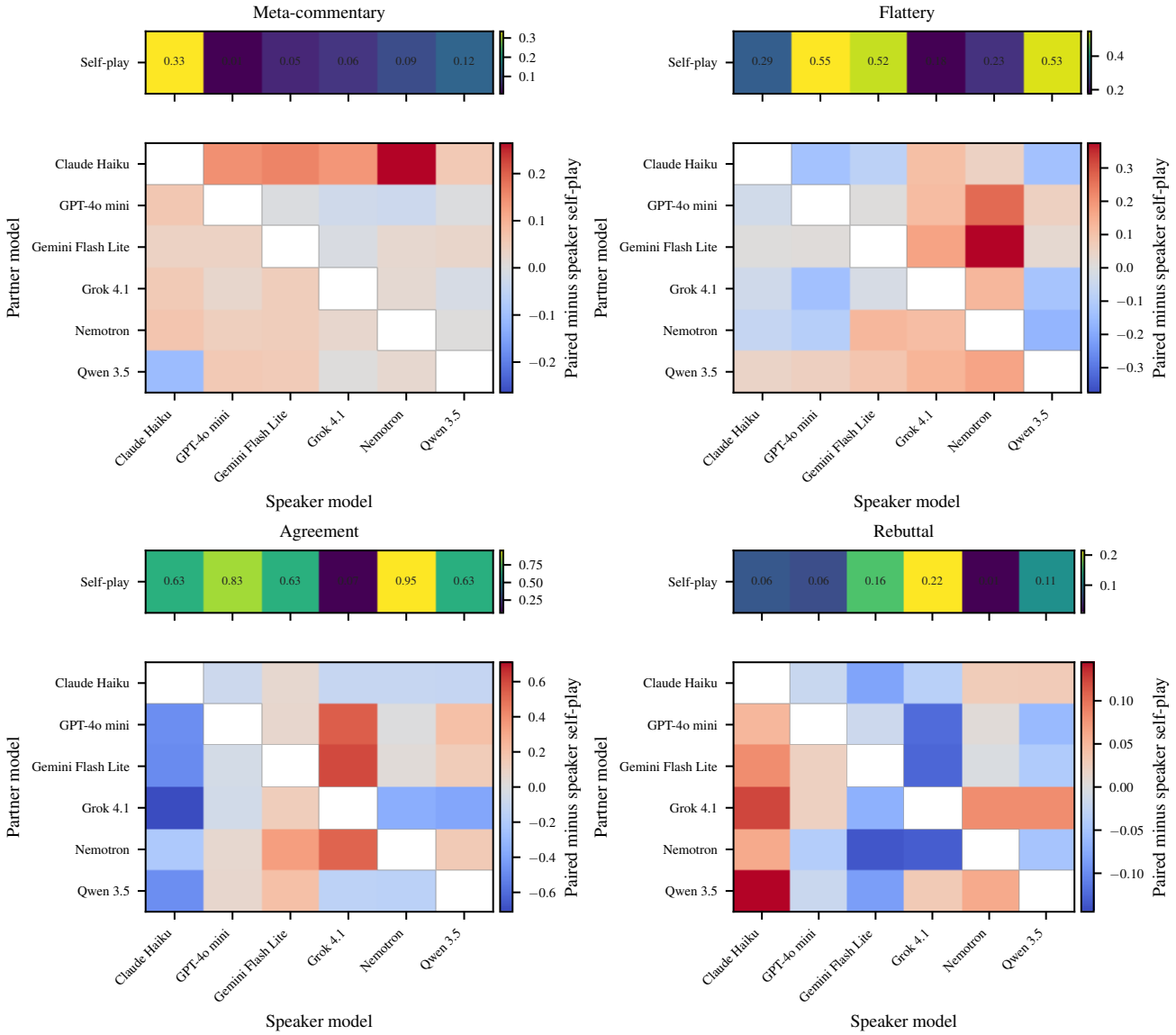


Figure 15. Influence-on-partners heatmaps across all turns. Rows index influencing models and columns index affected partner models. Each off-diagonal cell shows the affected partner minus self-play value for pairing affected partner S with influencing model M; row averages therefore recover the mean and delta influence summaries. Companion self-play calibration values belong to the affected partner columns, not the influencing-model rows.

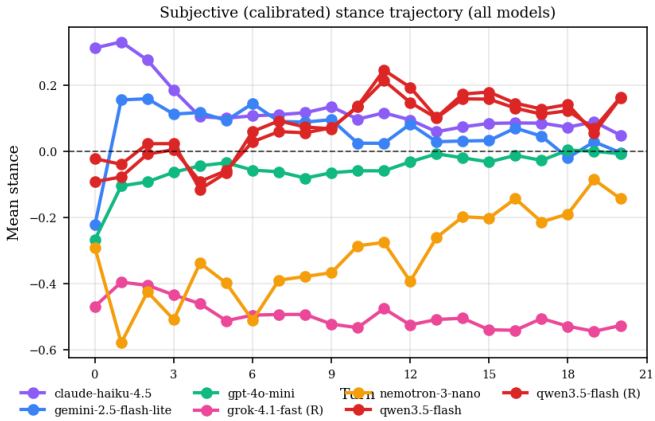


Figure 16. Mean stance trajectory, centered by their control settings.

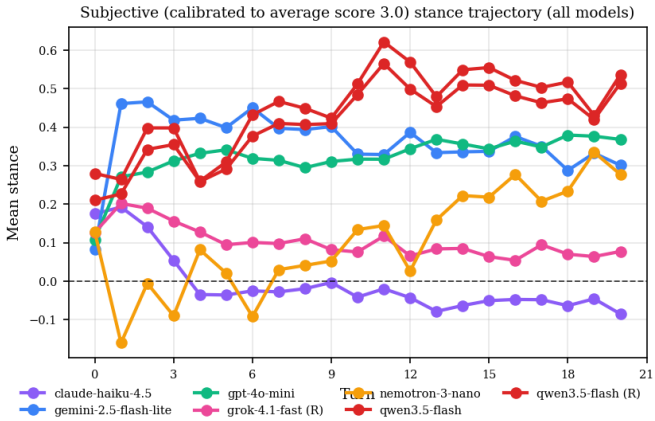
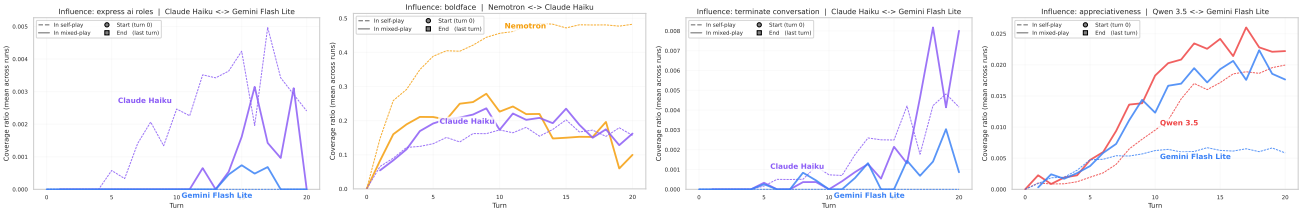


Figure 17. Mean stance trajectory, centered by the middle value within stance range.



(a) AI-role expression *(b)* Boldface *(c)* Conversation termination *(d)* Appreciativeness

Figure 18. Examples of feature-level influence in different model pairs. Solid lines show models' mixed-play behavior, and dashed lines show self-play behavior. **(a)–(c)** Claude Haiku shows a strong pull on explicit AI-role expression, boldface formatting, and conversation-termination language. Affected partner models shift toward Claude-associated behavior in these dimensions during interaction. **(d)** Appreciativeness shows Qwen pulling Gemini.

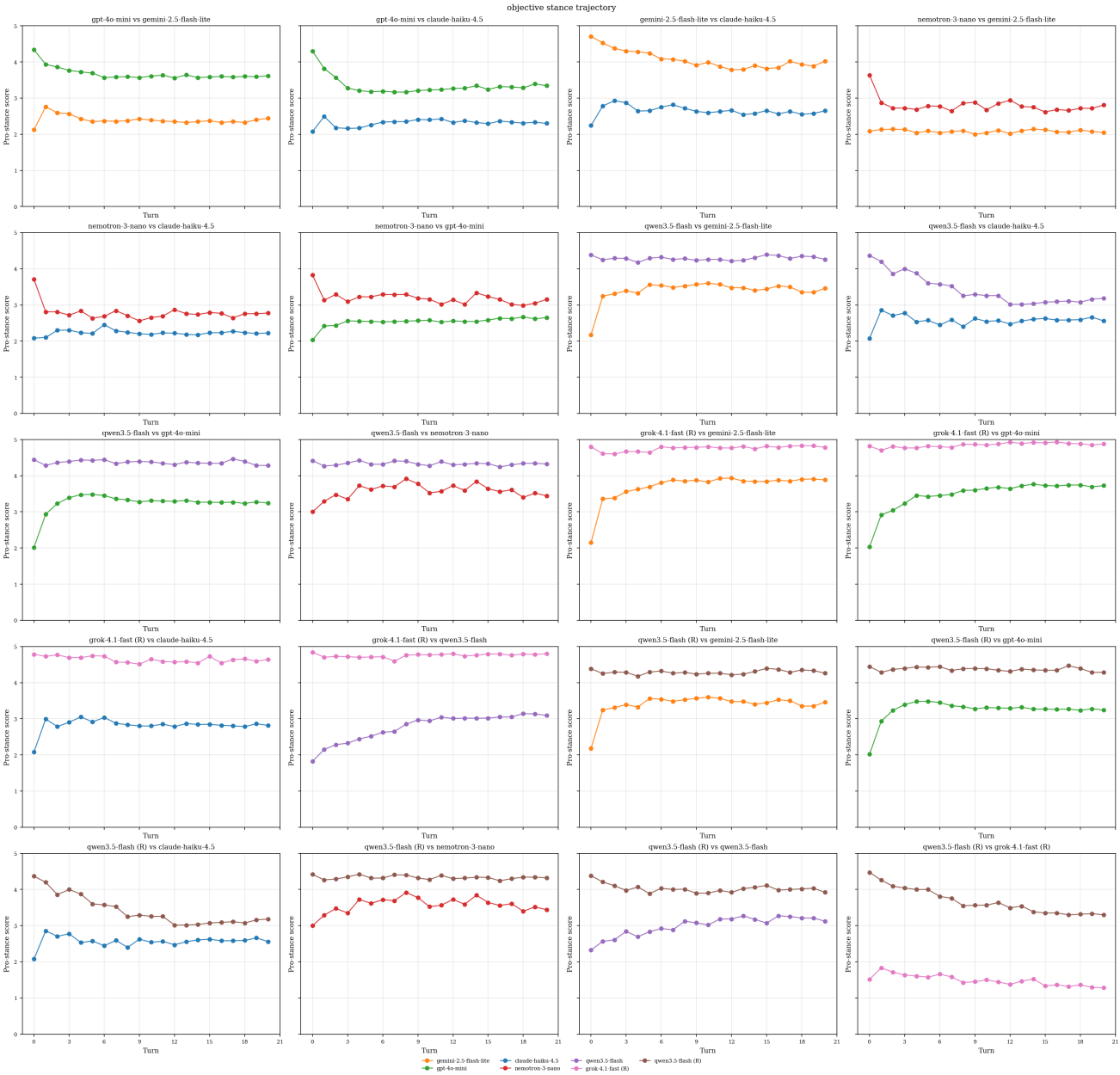
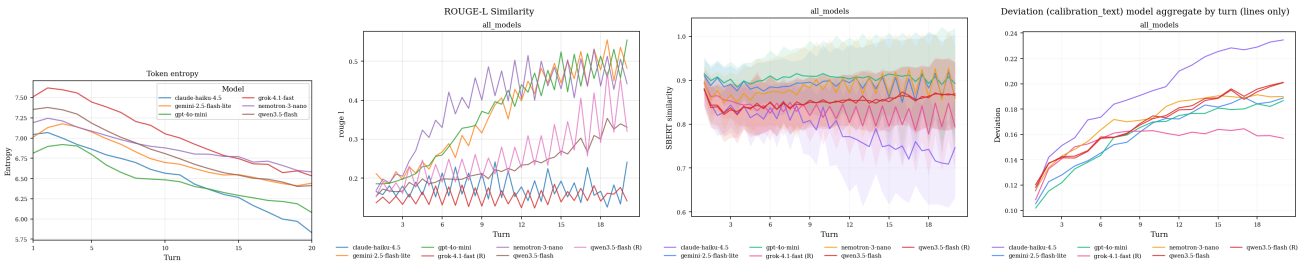


Figure 19. Pairwise stance trajectory.



(a) Lexicon entropy *(b)* ROUGE-L similarity *(c)* Semantic similarity *(d)* Topic deviation

Figure 20. **(a)–(c)** Lexicon entropy decreases and ROUGE-L similarity increases over turns, indicating lexical compression, while semantic similarity remains flat or decreases, indicating continued semantic diversity. **(d)** Conversations continue to drift away from the initial topic over time.