Attractor States Emerge in Multi-Turn LLM Conversations

Ting-Wen Ko [1 2] **Jonas Geiping** [1 2 3]

Abstract

Large language models (LLMs) are increasingly used in open-ended 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*, i.e. topic-independent stable sets of behaviors which conversations settle into. Across 7 LLMs and 20 controversial topics, we compare self-play and mixed-play dyadic debates, tracking trajectories in representation space, discourse traits, and stances. We find self-play trajectories to be model-specific attractors that draw their conversation partners asymmetrically in mixed-play debates, influencing the other models' stylistic choices and behavior. For example, Claude Haiku is a strong attractor of other models in latent space, corresponding to other models taking on its traits like metacommentary, and models like GPT-4.1 nano are especially malleable. Our results suggest that open-ended LLM interactions are partially predictable from model-specific attractors, but shaped by structured and asymmetric partner influence. Overall, our analysis sheds some light on the complex behavior of open-ended multi-agent interaction, which we hope is helpful in designing, predicting, and monitoring autonomous agentic systems in the real world.

1. Introduction

Large Language Models (LLMs) are nowadays not only used in task-specific scenarios (Thakkar et al., 2026; Hughes et al., 2026), but also increasingly for open-ended tasks like conducting research and policy making (Gottweis et al., 2026; Li et al., 2025). In these settings, models act as agents, generating outputs that become inputs to subsequent model calls, and pursuing broader goals over multiple turns.

[1]Max Planck Institute for Intelligent Systems [2]ELLIS Institute Tübingen [3]Tübingen AI Center. Correspondence to: Ting-Wen Ko <ting-wen.ko@tue.ellis.eu>.

Accepted at the Workshop on Trustworthy AI for Good (AI4GOOD) at the 43rd International Conference on Machine Learning (ICML), 2026. Copyright 2026 by the author(s).

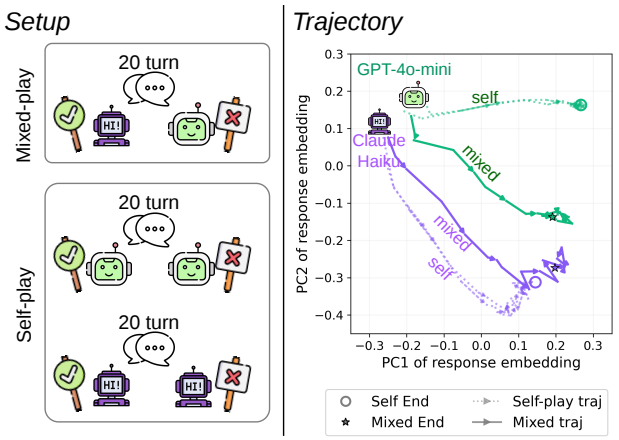


Figure 1. **Left.** We study 20-turn debate between two LLM-agents using two setups: 1) *mixed-play* where agents are instantiated from different models, and 2) *self-play* where agents are from the same model, which is also the control group that we later observe as "attractor" (Sec. 4.2). In both settings we assign one agent to be supporting and the other to be opposing a controversial topic. **Right.** The 20-turn mixed-play trajectory of Claude vs GPT and the respective self-play counterparts in the 2-d latent space spanned by the first two principal components of the topic-centered text embeddings of self-play responses. The two models' displacement from self-play to mixed-play are unequal, showing the asymmetrical attraction by different models' self-play attractors.

This multi-agent deployment in long-turn open-ended tasks raises several questions: limited diversity of generated content (Jiang et al., 2025a), shifts in language use (Abdulhai et al., 2026), amplified opinions (Nehring et al., 2024), and drift away from initial objectives (Laban et al., 2025; Ratnakar & Raghavendra, 2025). These effects arise gradually during the interactions between models, and it is hence critical to study their dynamics over the course of long-turn conversations.

However, relatively little is known about what actually happens when LLMs engage with one another over many turns. Most prior multi-agent LLM research focuses on settings with explicit objectives and measurable outcomes, such as math, coding, games, or factual reasoning, where success can be evaluated through utility, accuracy, or task completion metrics (Guzman Piedrahita et al., 2025; Du et al., 2024; Liang et al., 2024; Khan et al., 2024). Meanwhile, phenomena such as behavioral drift (Lu et al., 2026), mode collapse (Qiu et al., 2025), and being trapped by conversa-

tion history (Simhi et al., 2026) have mainly been studied in single-agent settings using simulated user responses or repeated generations. To better understand open-ended systems, anticipate behavior in settings with limited oversight, and prepare for increasingly autonomous multi-agent deployments, we focus on multi-turn LLM interaction itself. Specifically, we ask: *how are long-turn model-model interactions organized*? In particular, do conversations reveal stable model-specific regimes, and how do those regimes interact when different models are paired?

To study this, we run 20-turn debates between pairs of LLMs and construct pairwise combinations among seven open- and closed-source models. We analyze these interactions by separating them into *self-play*, where both participants are instances of the same model, and *mixed-play*, where different models interact. While we focus on mixed-play dynamics, self-play provides a baseline for understanding each model's intrinsic conversational behavior.

We find that self-play interactions reveal broad but reproducible model-specific endpoint attractors in the latent space. Using these attractors as anchors, we decompose interactions into quantitative measures along the axis connecting two corresponding self-play attractors. It turns out that mixed-play behavior is organized around the corresponding self-play attractors of both models. Also, model attractors show asymmetric influence: some models, like Claude Haiku, are relatively resistant to change while exerting strong influence on their partners, whereas others, such as GPT-4.1 nano, are more malleable. We further conduct a discourse-trait and stance analysis to match these behaviors in the latent space to concrete evidence of corresponding behavioral transfer in conversational style. Figure 1 provides an overview of the paper. The left panel illustrates the experimental setup, and the right panel shows a concrete example trajectory from one debate, showing both models moving toward each other in representational space.

Together, the findings suggest that multi-turn LLM interactions depend less on the capabilities of individual models, but also on which models are paired, which models are more influential, and which traits propagate through interaction. We hope this study serves as a framework to study the emergent behavior within the interaction of agents, and can be helpful in designing better tools to measure and predict agents' behavior in the foreseeable future of large systems of interconnected autonomous agents.

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 Kaur (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. (2025b); 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. (2025a;b) 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. Choi et al. (2025) finds that LLMs can infer characteristics of their conversational partners from reasoning patterns, linguistic style, and alignment preferences, and adapt their behavior accordingly.

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) construct

a proof 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.

Yet, Anthropic (2025) also describe a "spiritual bliss" attractor state in Claude Opus 4 self-interactions 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 the possibility of *persona drift*: the tendency of models to gradually move from their provider-specified personality over the course of an 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: Measuring LLM-LLM Dynamics

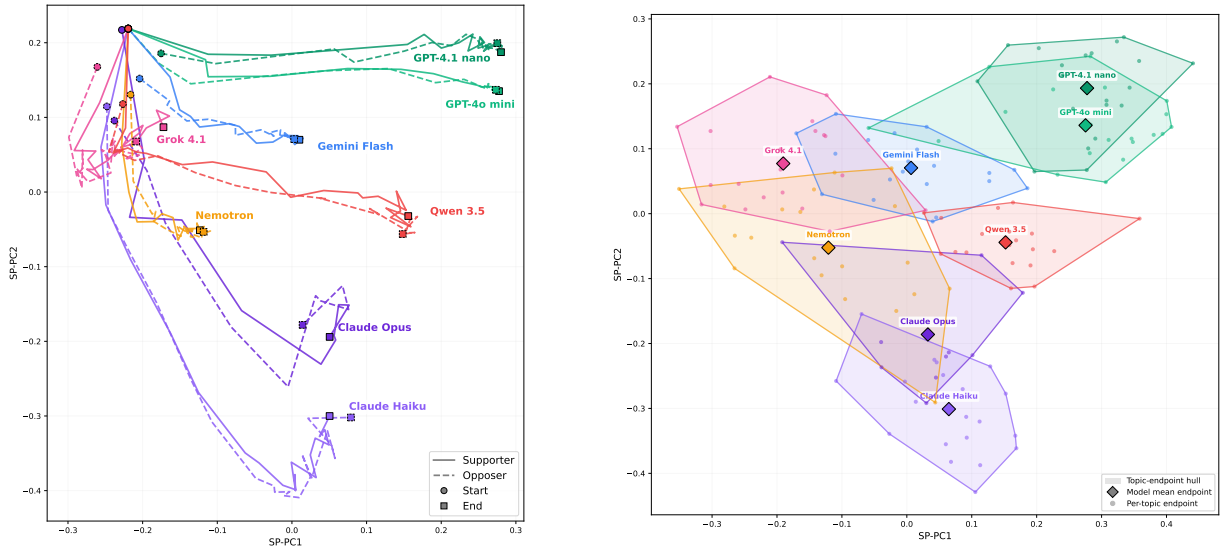
Task. We study interaction dynamics through open-ended debates between two LLM agents. Unlike tasks with fixed objectives, controversial debates allow agents to frame issues, shift stances, and develop shared conversational norms over multiple turns. This setting lets us examine whether one model's characteristic behavior pulls another model toward its style, framing, or stance, a dynamic that is important for understanding autonomous multi-agent deployments under limited human oversight.

We pick 20 controversial social and policy topics, and pair each topic with three pro and con reference statements from ProCon.org (App. C). In each run, we start the conversation with a neutral topic-specific statement. Then, two LLM agents discuss that topic for $T = 20$ turns. Each run thus produces one trajectory of model responses.

Main Experimental Conditions. Our main experiments compare self-play and mixed-play debate. In the **self-play debate** condition, both agents are independent instances of the same model and are assigned opposing roles, SUPPORTER/OPPOSER. This condition provides a same-model baseline for where each model tends to end up under the debate protocol. In the **mixed-play debate** condition, the two agents are instantiated from distinct models while retaining the same role assignment. Comparing mixed-play against self-play isolates cross-model interaction effects under a fixed debate protocol, allowing us to test whether one model moves another toward its self-play endpoint region and whether this movement is asymmetric across model pairs.

Ablations. We also include a **self-play stance-free ablation**, in which both agents are independent instances of the same model and are assigned the neutral role DISCUSSANT. Comparing this stance-free condition with self-play debate helps separate model-specific multi-turn drift from dynamics induced by the SUPPORTER/OPPOSER role configuration.

Implementation. In the debate conditions, agents receive in system prompt role-specific instructions and corresponding



(a) Trajectories of self-play responses.

(b) Self-play endpoint hulls.

Figure 2. **Self-play mean trajectories and endpoints. (a)** Self-play mean trajectories separate over turns **(b)** Self-play endpoint basins occupy broad, model-specific regions across topics in the latent space, here shown by PCs of topic-centered embedding of all turns.

reasons to support or oppose the statement; in the stance-free ablation, both agents receive the neutral DISCUSSANT instruction. This information-symmetric design makes results easier to attribute to model identity, role configuration, and interaction regime rather than to asymmetric access to evidence. (App. C.2.)

Our agents are instantiated from the following 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.[1]

4. Latent Interaction Dynamics

We now analyze the resulting conversations as trajectories in embedding space. Self-play provides the reference geometry: it tells us where each model tends to end up under the fixed debate protocol. Mixed-play is then analyzed relative to this reference, allowing us to ask whether paired models move closer together (Sec. 4.2), whether this movement is symmetric across model pairs (Sec. 4.3), and whether their endpoints lie along the axis connecting their self-play regions (Sec. 4.4, 4.5, 4.6).

4.1. Latent-space setup

We embed each response as a 384-dimensional SBERT vector e_i (Reimers & Gurevych, 2019). To remove topic-level offsets while retaining variation associated with model, role, and interaction condition, we center embeddings within topic:

$$x_i = e_i - \frac{1}{|\mathcal{I}_k|} \sum_{j \in \mathcal{I}_k} e_j, \tag{1}$$

where $\mathcal{I}_k$ denotes the set of responses for topic k. We compute principal components (PCs) using only the topic-centered self-play embeddings; we refer to these axes as SP-PCs. We use this representation in two ways. For visualization, we project both self-play and mixed-play trajectories onto the first two SP-PCs, so that all trajectories are displayed in a common self-play reference basis. For quantitative analyses, we use the full 384-d topic-centered embeddings x_i.

We also analyze endpoints i.e. the final-turn response. For model A and topic k, let $s_{A,k}$ denote the topic-centered self-play endpoint embedding. For a mixed-play pair (A, B), let $m_{A|B,k}$ denote model A's topic-centered endpoint embedding when paired with model B. For all metrics, we compute numbers per topic before aggregation into final statistics unless otherwise stated.

4.2. Self-play defines attractor-like model basins

We first characterize self-play as a latent-space baseline. Specifically, we ask whether each model's late-turn responses settle into bounded, reproducible endpoint regions rather than diffusing indefinitely or collapsing into a shared cross-model region. We refer to these regions as *attractor-like basins* in analogy to dynamical-systems attractors (Strogatz, 2018), used here to describe model-specific stable endpoint regimes.

Metrics. To investigate this, we project trajectories and end-

[1] CLAUDE-4.5-OPUS is run only in 10-turn self-play conversations due to budget constraints.

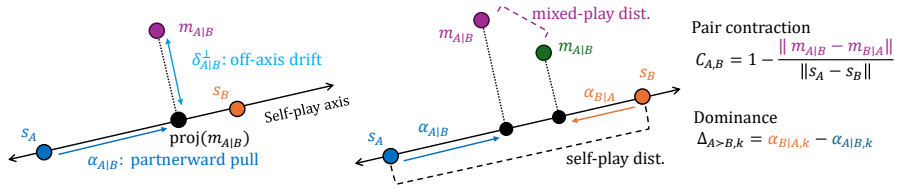


Figure 3. **Mixed-play endpoint metrics. Left:** $m_{A|B}$, model A's endpoint when paired with model B, is decomposed relative to the self-play axis from self-play endpoints s_A to s_B. Partnerward pull $\alpha_{A|B}$ measures interpolation along this axis, while off-axis drift $\delta^{\perp}_{A|B}$ measures displacement not explained by one-dimensional consensus. **Right:** Pair contraction C_{AB} measures how much closer the two mixed-play endpoints become relative to their self-play separation, while dominance $\Delta_{A \succ B}$ measures asymmetry in directional pull.

Table 1. Nearest-cluster separation endpoint statistics in the full 384-D topic-centered embedding space, using self-play endpoints with roles averaged. S_{basin} is the set-to-set basin separation score. Full endpoint diagnostics, including the endpoint F-ratio, within-model spread, S_{cent}, and nearest-cluster models, are in App. D.

	Gemini Flash	Qwen 3.5	GPT-4o mini	Claude Haiku	Grok 4.1	Nemotron	GPT-4.1 nano	Claude Opus
S_{basin}	4.08	2.39	2.05	2.31	1.55	1.50	1.81	1.84

points onto the first two SP-PCs as described in Section 4.1. To assess local cluster separation, we compute a basin separation score S_{basin} as follows. This statistic compares a model's within-cluster endpoint spread against its pairwise distance to the closest model's endpoint cluster. Specifically, for each model A, let

$$\mu^s_A = \frac{1}{n_A} \sum_{j=1}^{n_A} s_{A,j} \tag{2}$$

be the centroid of this endpoint set. We define the within-cluster spread as

$$W_s(A) = \frac{1}{n_A} \sum_{j=1}^{n_A} \|s_{A,j} - \mu^s_A\|^2 . \tag{3}$$

Then, the pairwise set-to-set distance between model A and another model $B \neq A$ is

$$d^2_{\text{set}}(A, B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \|s_{A,i} - s_{B,j}\|^2 . \tag{4}$$

Finally, the basin separation score is defined as

$$S_{\text{basin}}(A) = \frac{\min_{B \neq A} d^2_{\text{set}}(A, B)}{W_s(A)}. \tag{5}$$

Values above one mean that the closest endpoint set is farther away than the model's own endpoint spread. More endpoint analysis is provided in App. D.

Results. Fig. 2a shows that the self-play trajectories begin from a shared region but progressively separate over turns, moving toward its own model-specific characteristic endpoint regime. This indicates that late-turn behavior is not simply a continuation of the common prompt initialization. Fig. 2b shows the same pattern at the endpoint level. Endpoints do not intermingle diffusely across models, but occupy locally separated model-specific regions. We interpret these regions as *attractor-like basins*: bounded endpoint regimes that repeatedly arise under a given model's self-play dynamics.

Table 1 also shows the basin separation score S_{basin} exceeding one for every model, indicating that each model endpoint basin remains locally separated from its closest model. Thus, although each basin has non-negligible internal variability across topics, its variability remains smaller than its separation from the nearest basin.

Role and seed ablations. We run two ablations (Fig. 4) to check these basins are artifacts of the debate role setting or a particular random seed. First, replacing Supporter/Opposer role-play with neutral DISCUSSANT/DISCUSSANT self-play preserves the same qualitative pattern: trajectories initialized from the same opening still separate into model-specific endpoint basins. This suggests that the observed basins are not merely induced by adversarial role-play. Second, repeating selected self-play settings over three seeds also returns to comparable endpoint basins. Together, these controls support the interpretation that self-play exposes model-intrinsic conversational basins: recurrent endpoint regimes shaped primarily by the generating model, while still allowing within-basin variation across topics, roles, and seeds. A similar pattern under an alternative embedding is provided in App. D.6.

4.3. Mixed-play reveals basin-to-basin attraction

Building on the basis of self-play model basins, we then turn to mixed-play debate setting to see how different models interact. Visually, we immediately observe in Fig. 5 that mixed-play trajectories seem to shift away from a model's self-play centroid toward its partner's region. But do two

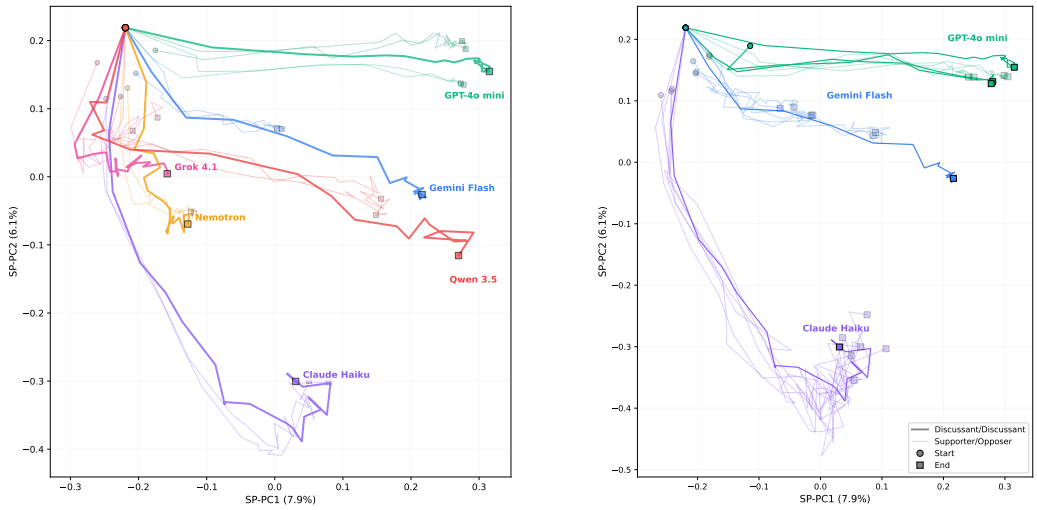


Figure 4. **Self-play basins are stable. Left**: neutral DISCUSSANT/DISCUSSANT self-play trajectories projected onto the self-play PCA space. Even without pro/con roles, trajectories separate into model-specific regions. **Right**: selected Supporter/Opposer self-play settings repeated across three seeds, where each line shows a repeated run. We can see that repeated runs return to comparable endpoint regions.

models come closer after interaction? If so, how close?

Metric. We answer this by computing *pair contraction* for each unordered agent pair (A, B) as

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

where positive C means the two mixed-play endpoints are closer to each other than the corresponding self-play endpoints, indicating consensus-like contraction (refer to Figure 3 for a schematic). In Table 2, 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 vs GPT-4.1 nano (68.0%) and GPT-4o mini vs Gemini Flash (52.7%), but is weak for several Claude/Grok/Nemotron pairs and slightly negative for Grok 4.1 vs Gemini Flash (−1.3%). Therefore, agents do not always converge after interaction and this differ from pair to pair, but they do generally show partial consensus.

4.4. Mixed-play displacement is asymmetric

The preceding analysis shows that mixed-play endpoints contract relative to the corresponding self-play endpoints. However, contraction alone does not specify *where* each model moves. Two endpoints can become closer because both models move equally toward each other, because one model is pulled more strongly toward the other, or because both endpoints drift in a shared direction outside the line connecting their self-play basins. We therefore decompose mixed-play displacement relative to the self-play axis between the two models.

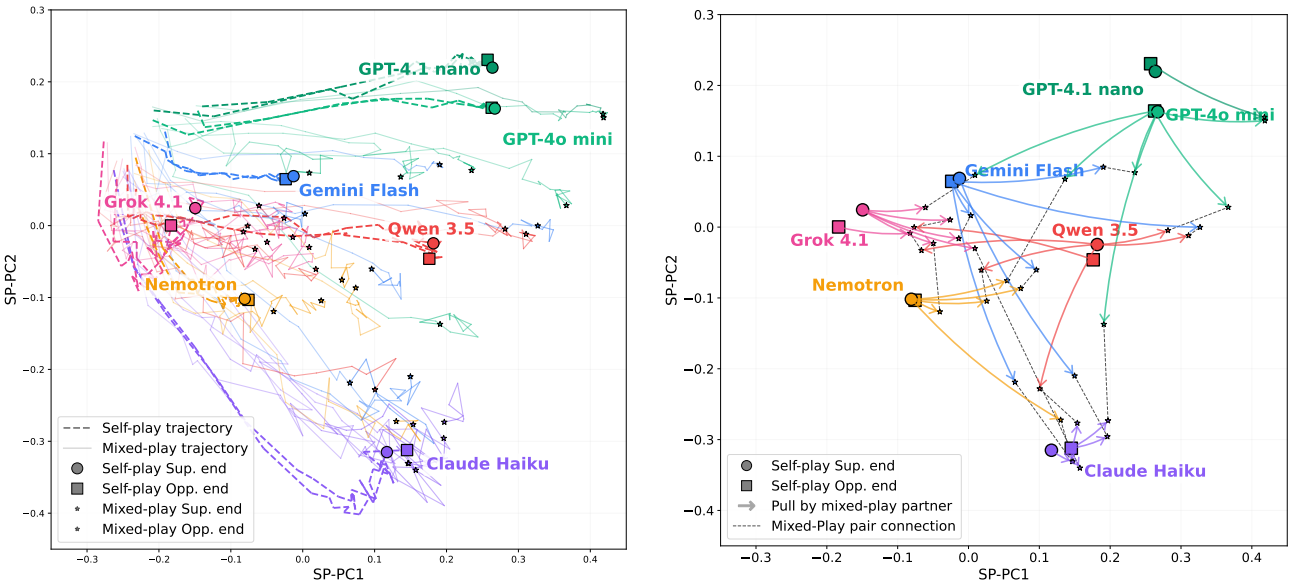
This decomposition is useful because self-play gives a

Table 2. Pair-level dominance and contraction. $\Delta_{A\succ B} > 0$ means the first model pulls the second more than it is pulled; $\Delta_{A\succ B} < 0$ means the vice versa. $C_{A,B}$ is pair contraction. Bootstrap confidence intervals are reported in App. D.3.1.

Pair	$\Delta_{A\succ B}$	$C_{A,B}$ (%)
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

model-specific reference for what each agent tends to become without cross-model interaction. For a given topic, the line from $s_{A,k}$ to $s_{B,k}$ represents the direction from model A's self-play behavior toward model B's self-play behavior. Movement along this line captures partner-aligned attraction: whether one model is pulled toward the other's intrinsic endpoint basin. Movement orthogonal to this line captures residual displacement not explained by such interpolation. Thus, the decomposition answers two questions: 1) *who pulls whom toward whose self-play basin?* (Sec. 4.4) 2) *is that partnerward pull sufficient to explain where the endpoint lands?* (Sec. 4.6). In this subsection, we focus on the first question.

Metric. Our decomposition (Fig. 3) separates displace-



(a) Self- and Mixed-Play Trajectories. *(b)* Endpoint pull.

Figure 5. **Models attract each other in mixed play**. **(a)** We overlay mixed-play trajectories on the self-play trajectories in Fig. 2a. Solid lines and small markers are mixed-play trajectories and endpoints; dashed lines and large points are self-play ones. **(b)** Focusing only on endpoints, we highlight the mixed-play centroids pulled away from their self-play counterparts.

ment into 1) partnerward interpolation and 2) off-axis drift[2]. Specifically, for mixed-play between agent pair (A, B) and topic k, we first define the self-play axis

$$v_{A\to B,k} = s_{B,k} - s_{A,k}. \tag{7}$$

as the local spectrum between the two self-play basins. Then, the *partnerward pull* of model A when paired with model B is

$$\alpha_{A|B,k} = \frac{(m_{A|B,k} - s_{A,k})^\top v_{A\to B,k}}{\|v_{A\to B,k}\|^2}. \tag{8}$$

Values $0 < \alpha < 1$ indicate interpolation from A's self-play endpoint toward B's endpoint, whereas larger values indicate stronger pull toward its partner model B's basin. Negative values indicate movement away from the partner; and values above one indicate overshoot beyond that.

We also derive the *dominance* of A over B as

$$\Delta_{A\succ B,k} = \alpha_{B|A,k} - \alpha_{A|B,k}. \tag{9}$$

Positive values mean the mixed-play pair is directionally skewed toward A's self-play endpoint, as B moves more toward A than A moves toward B.

Results. The dominance scores in Table 2 show that mixed-play influence is strongly pair-specific and non-symmetric. The largest absolute scores indicate Claude Haiku's directional dominance over Nemotron ($\Delta = -0.538$), GPT-4o mini ($\Delta = -0.507$), Gemini Flash ($\Delta = -0.484$); and Qwen 3.5 over Gemini Flash ($\Delta = 0.396$). These patterns show that some models exert stronger directional pull over their partners, and the direction and magnitude of this pull depend on the specific model pair. Bootstrap confidence intervals are reported in App. D.3.1, Table 8.

[2]We average endpoints per topic to compute the decomposition to avoid overstating mean contraction which could happen if we average over centroids instead (App. D.4.)

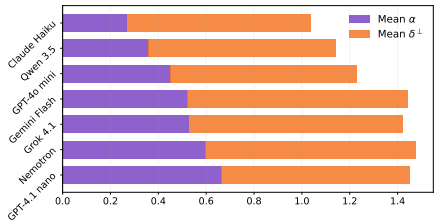


Figure 6. **Mixed-play endpoint decomposition.** Purple and orange bars denote the partnerward pull α and normalized off-axis drift $\delta^\perp$ respectively, along the same-topic self-play axis. Most endpoints lie in the interpolation regime $0 < \alpha < 1$, but they also retain nonzero off-axis displacement $\delta^\perp$.

4.5. Models differ in attraction and malleability

The pair-level dominance results show that mixed-play influence is asymmetric. We next ask whether these pairwise asymmetries reflect broader model-level profiles: *which models are generally more movable, and which behave more like stable attractors for their partners*. To answer this, we aggregate the partner-aligned component of the endpoint decomposition by model.

Results. Fig. 6 shows that most endpoints fall in the interpolation regime $0 < \alpha < 1$, meaning that a model's endpoint

usually moves between its own same-topic self-play endpoint toward its partner's. Aggregated by model, GPT-4.1 nano has the largest partnerward pull ($\alpha = 0.665$), followed by Nemotron (0.588), and Gemini Flash (0.540). Claude Haiku has the lowest pull (0.266), indicating the strongest resistance to partnerward displacement.

Together with the dominance scores in Table 2, these model-level averages suggest distinct attraction–malleability profiles. Claude Haiku is the clearest resistant attractor: it is the least partnerward-pulled model overall and directionally dominates all partners. GPT-4.1 nano shows the opposite profile: it has the largest average partnerward pull, suggesting high malleability rather than directional dominance.

Regarding the first decomposition question: *who pulls whom toward whose self-play basin*, we now see that mixed-play interaction is not a uniform averaging process. Some models are more movable, some are more resistant, and some exert influence selectively depending on the partner.

4.6. Mixed-play is not one-dimensional averaging

However, the partnerward pull α captures only one interpretable mode of cross-model influence: movement from a model's own self-play endpoint toward its partner's self-play endpoint. Mixed-play may also change endpoints in directions that are not aligned with this partnerward axis. The off-axis component therefore addresses the second part of the decomposition question: *is partnerward pull sufficient to explain where the endpoint lands?*

Metric. We define *normalized off-axis drift* as

$$\delta^{\perp}_{A|B,k} = \frac{\left\|(m_{A|B,k} - s_{A,k}) - \alpha_{A|B,k} v_{A\to B,k}\right\|}{\|v_{A\to B,k}\|}. \quad (10)$$

This quantity captures displacement orthogonal to the self-play axis, i.e., movement not explained by one-dimensional interpolation between the two self-play endpoints. Again, Figure 3 provides a schematic for this metric.

Results. Model-level raw off-axis drifts $\delta^{\perp}$ are relatively concentrated around 0.8. (Fig. 6): Grok 4.1 (0.888), Nemotron (0.871), and Gemini Flash (0.869) have the largest off-axis residuals, while GPT-4.1 nano (0.788), GPT-4o mini (0.773), Qwen 3.5 (0.773), and Claude Haiku (0.734) are lower. Thus, the self-play axis captures the main partnerward direction, but not the whole displacement.

Fig. 6 thus gives a two-part view of mixed-play. First, endpoints mostly interpolate between same-topic self-play basins. Second, endpoints retain a substantial off-axis component, showing that mixed-play is not simply one-dimensional averaging between two model centroids.

4.7. Summary

To sum up, the results in this section first show that self-play reveals broad model-specific basins. Then, mixed-play usually moves endpoints partnerward along the same-topic axis between those basins, but endpoints also retain raw off-axis displacement and the amount of pull depends 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. GPT-4.1 nano in contrast is highly partnerward-pulled ($\alpha = 0.665$) and contracts strongly with GPT-4o mini, suggesting high malleability.

5. What Behaviors do these Attractors Correspond to?

So far, we have looked at geometric analyses showing how conversations move in the latent space. We now ask what those endpoint regions and attraction effects correspond to behaviorally. We first consider eight **conversational traits**, such as *agreement*. We then examine how agents' **stance** toward the topic changes across turns: whether initially opposed agents remain polarized, converge toward neutrality, or move toward one side.

5.1. Conversational Traits

Setup. We use a GPT-OSS-20B as judge to score turn-level discourse traits with scale 0–1 after the main experiment. We report model-level means and influence for eight selected traits. See detailed trait definitions in Table 15.

Results: Model Signature. For model A and trait f, let $f_A^{s,(t)}$ denote the turn-level judge score in self-play. We first compute the model-level self-play mean

$$\bar{f}_A^s = \frac{1}{T}\sum_{t=1}^{T} f_A^{s,(t)}, \quad (11)$$

over T turns. Fig. 7 shows the distinct behavioral profiles among different models during conversations. Claude Haiku is the model with the most meta-commentary (0.331), well above Qwen 3.5 (0.116) and others. This aligns with case studies in which Claude comments on the structure or constraints of the conversation itself, e.g., "The conversation was real and constrained." (App. E.7)

Others occupy different behavioral regions. Grok 4.1 remains the most adversarial and rational, 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), low rebuttal (0.028), and high agreement (0.837), which seems like confident convergence. Full trait definition and result tables are in Tables 18, 19, 20, 21, with more discussion and case studies in App. E.

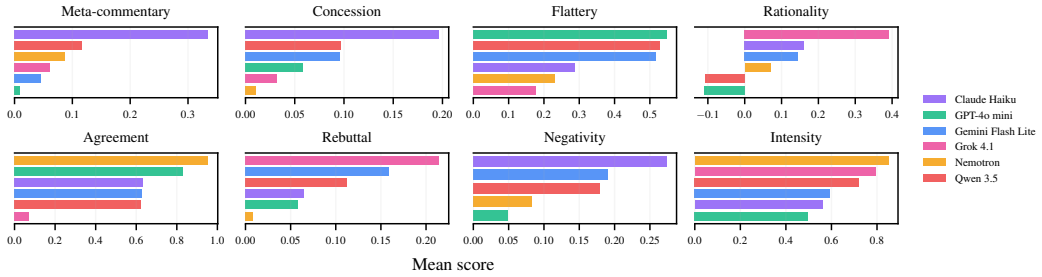


Figure 7. **Model-specific self-play discourse signatures**. Claude Haiku stands out on meta-commentary, while other models differ in flattery, rationality, agreement, rebuttal, negativity, and intensity. Complete trait tables are provided in App. E.

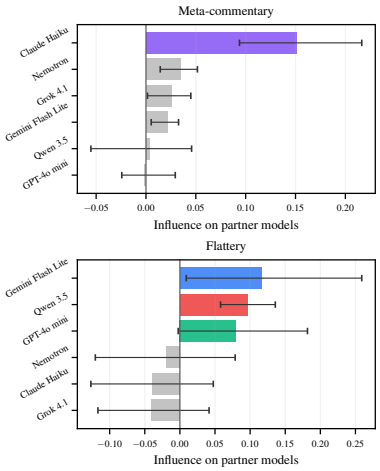


Figure 8. **Trait-level partner influence in mixed-play.** Claude Haiku most strongly pulls its partners toward meta-commentary, while Gemini Flash Lite, GPT-4o mini, and Qwen 3.5 most strongly pull their partners toward flattery.

Results: Trait Transfer. For model A and partner B, let $f^{m,(t)}_{B|A}$ denote the turn-level judge score of the *affected* model B in mixed-play when paired with A. We define the partner-conditioned trait transfer as

$$\tau^{f}_{B\leftarrow A} = \frac{1}{T}\sum_{t=1}^{T} f^{m,(t)}_{B|A} - \frac{1}{T}\sum_{t=1}^{T} f^{s,(t)}_{B}, \qquad (12)$$

where positive values indicate that B expresses trait f more when paired with A than in its own self-play. The outgoing trait influence of model A is then

$$\bar{\tau}^{f}_{A} = \frac{1}{|\mathcal{M}\setminus\{A\}|}\sum_{B\in\mathcal{M}\setminus\{A\}} \tau^{f}_{B\leftarrow A}, \qquad (13)$$

where $\mathcal{M}$ is the set of all models.

In Fig. 8, Claude Haiku acts as a meta-commentary attractor: partners paired with Claude show the largest increase in meta-commentary, making conversation reflective about the interaction process. By contrast, flattery—praise of the partner or of the conversation itself—is most strongly induced by Gemini Flash Lite, GPT-4o mini, and Qwen 3.5. Claude already exhibits elevated meta-commentary in self-play, and this trait transfers in mixed-play; similarly, models with stronger self-play flattery signatures tend to pull their partners toward more flattering discourse. That is, the latent-space interactions (Sec. 4) are manifested in behavior through trait-level transfer. We leave the discussion of how stances change in Appendix A.

6. Conclusion

We study open-ended model–model discussions and find that their trajectories are structured by model-specific attractor-like basins. In self-play, discussions tend to settle into reproducible endpoint regions associated with each model. In mixed-play, endpoints often move along the axis connecting the two models' self-play basins, i.e. models in mixed play end up influencing each others behaviors, which shows up as a move toward each other in representation space. This influence is asymmetric: some models, such as Claude Haiku, appear more resistant, i.e. moving little themselves, and attract partners more strongly to their set of traits, such as meta-commentary, whereas others, such as GPT-4o mini, are more malleable, being easily movable by other models' traits. We link the geometric dynamics to asymmetric transfer of concrete discourse traits, such as meta-commentary and flattery. In terms of stance dynamics, these attractors are not reducible to convergence: self-play does not consistently amplify stance, and mixed-play does not always produce compromise.

Together, these findings suggest that multi-agent LLM interactions remain structured even in open-ended settings. Their dynamics are shaped by model-specific basins, asymmetric influence, and interaction-specific behavioral transfer. We hope this work encourages more systematic study of LLM interaction dynamics beyond debate, including the development of formal tools for characterizing how agentic systems *influence, resist*, and *reshape* one another in deployment.

Acknowledgments

This research was partially supported by the EC under the grant No. 101195233 (OpenEuroLLM). Jonas Geiping acknowledges the support of the Hector foundation. This research received support through Schmidt Sciences within

the project long-term safety behavior of LLM-based agents. The authors thank Shashwat Goel, Guinan Su, Sajad Movahedi, Xueyan Li, Changling Li, Johannes Zenn, Hsun-Yu Kuo, and Patrik Wolf for valuable and constructive discussions.

References

Abdulhai, M., White, I., Wan, Y., Qureshi, I., Leibo, J., Kleiman-Weiner, M., and Jaques, N. How LLMs Distort Our Written Language, March 2026. URL http://arxiv.org/abs/2603.18161. arXiv:2603.18161 [cs.CL].

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.

Choi, Y., Li, C., Yang, Y., and Jin, Z. Agent-to-agent theory of mind: Testing interlocutor awareness among large language models. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28895–28928, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1471. URL https://aclanthology.org/2025.emnlp-main.1471/.

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.

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/.

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Sirkovic, P., Myaskovsky, A., Glowaty, G., Weissenberger, F., Orlandi, A., Popovici, D., et al. Accelerating scientific discovery with co-scientist. *Nature*, pp. 1–3, 2026.

Guzman Piedrahita, D., Yang, Y., Sachan, M., Ramponi, G., Schölkopf, B., and Jin, Z. Corrupted by Reasoning: Reasoning Language Models Become Free-Riders in Public Goods Games, June 2025. URL https://ui.adsabs.harvard.edu/abs/2025arXiv250623276G. ADS Bibcode: 2025arXiv250623276G.

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, L., Chai, Y., Li, M., Liu, M., Fok, R., Dziri, N., Tsvetkov, Y., Sap, M., Albalak, A., and Choi, Y. Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond), October 2025a. URL http://arxiv.org/abs/2510.22954. arXiv:2510.22954 [cs].

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*, 2025b. URL https://arxiv.org/abs/2510.20039.

Kaur, A. Echoes of Agreement: Argument Driven Sycophancy in Large Language models. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22803–22812, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1241. URL https://aclanthology.org/2025.findings-emnlp.1241/.

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.

Laban, P., Hayashi, H., Zhou, Y., and Neville, J. LLMs Get Lost In Multi-Turn Conversation, May 2025. URL http://arxiv.org/abs/2505.06120. arXiv:2505.06120 [cs].

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.

Li, Y., Shen, X., Miao, Y., Yao, X., Ding, X., Krishnan, R., and Padman, R. Beyond single-turn: A survey on multi-turn interactions with large language models. *arXiv preprint arXiv:2504.04717*, 2025.

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.

Nehring, J., Gabryszak, A., Jürgens, P., Burchardt, A., Schaffer, S., Spielkamp, M., and Stark, B. Large Language Models Are Echo Chambers. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 10117–10123, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.884/.

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.

Qiu, T. A., He, Z., Chugh, T., and Kleiman-Weiner, M. The Lock-in Hypothesis: Stagnation by Algorithm, June 2025. URL http://arxiv.org/abs/2506.06166. arXiv:2506.06166 [cs].

Ratnakar, S. and Raghavendra, S. The Chameleon Nature of LLMs: Quantifying Multi-Turn Stance Instability in Search-Enabled Language Models, October 2025. URL http://arxiv.org/abs/2510.16712. arXiv:2510.16712 [cs].

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.

Simhi, A., Barez, F., Tutek, M., Belinkov, Y., and Cohen, S. B. Old Habits Die Hard: How Conversational History Geometrically Traps LLMs, February 2026. URL http://arxiv.org/abs/2603.03308. arXiv:2603.03308 [cs.CL].

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

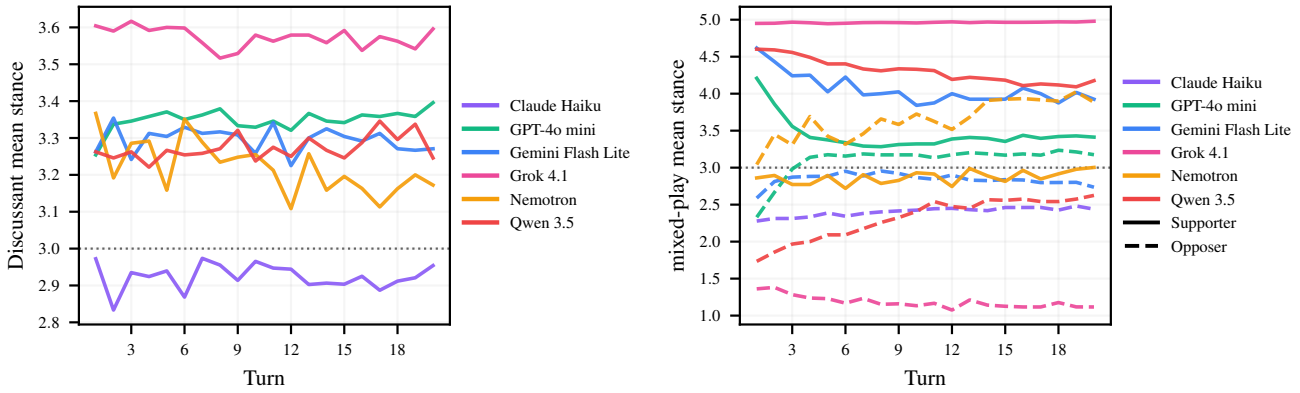
Tacheny, N. Geometric Dynamics of Agentic Loops in Large Language Models, January 2026. URL http://arxiv.org/abs/2512.10350. arXiv:2512.10350 [cs].

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.

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.



(a) Discussant stance baselines without assigned roles.

(b) Mixed-play stance, separated by model and assigned roles.

Figure 9. **Stance Changes during Debate. Left:** In the Discussant/Discussant setting, where no explicit Supporter/Opposer roles are assigned, models stabilize at different intrinsic stance levels. The full range of the stance is 1-5. **Right:** In mixed-play, stance trajectories do not follow a single convergence pattern across model pairs. Gemini, GPT, and Qwen tend to move toward weaker, more neutral stances in both Supporter and Opposer roles. Other cases show qualitatively different dynamics, including role reversal, such as GPT Opposers becoming more supportive, and strong persistence, such as Grok maintaining an extreme stance throughout the interaction.

A. Stance Changes in Discussion

Setup. We measure stance after each turn with a fixed Likert-scale questionnaire answered by the agent itself, conditioned on the full conversation history up to that turn. This design tests the models' self-reporting stance given the same condition as when they are actually generating the response. We discuss more on the setting, the questionnaire prompt, and computation details in App. C.2.

Results. First, we examine agents' stance trajectories in the self-play Discussant/Discussant setup (Fig. 9a). As a model-intrinsic baseline, agents here interact without any user-assigned stance to support/oppose the topic. It is worth noting that each model stabilizes on different stance scales instead of being neutral: Grok 4.1 remains more supportive, whereas Claude Haiku stays below the midpoint of 3.0.

As for the mixed-play Supporter/Opposer setting (Fig. 9b), we see that stance dynamics are not governed by a single convergence pattern. Some models show weakened stance intensity in both roles: Gemini, GPT, and Qwen tend to move closer to neutrality whether they are assigned as Supporter or Opposer. This suggests that, for these models, cross-model interaction can dampen the role-assigned stance rather than simply preserving the initial argumentative position. However, this neutralization pattern is not universal. In some cases, the assigned stance changes direction, as when GPT in the Opposer role becomes more supportive over the course of the dialogue. In other cases, the assigned stance remains highly stable, as with Grok, which maintains a strongly supportive stance throughout the interaction. [3]

Taken together, these results show that stances are only an incomplete view into the complex model-model dynamics analyzed through representation analysis and behavioral traits. In the Discussant/Discussant baseline, models stabilize at model-specific stance levels, but they do not simply reinforce their initial tendencies into increasingly extreme positions. Conversely, in the mixed-play Supporter/Opposer setting, explicit role assignment does not reliably produce convergence or compromise: some trajectories soften toward neutrality, whereas others reverse direction or preserve a strong assigned stance.

B. LLM Usage Declaration

This work used LLM assistance for literature review, coding support, and text polishing. All experimental design choices, analyses, interpretations, and final manuscript decisions were made by the authors.

[3] One may wonder the connection of the stance behavior with the close Supporter/Opposer endpoint regions observed in the latent space (Fig. 5a). We think the attractor-like basins appear to reflect broader model-specific discourse modes, only some of which are captured by explicit stance movement. For Gemini and GPT, latent proximity coincides with stance softening or consensus-like language, whereas for Grok, trajectories become representationally similar while remaining stance-separated.

C. Method Details

This appendix follows the organization of the main paper. We first collect methodological details, then report supplementary geometric checks for the endpoint-basin analysis, followed by behavioral and stance diagnostics used to interpret the basins.

C.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.

C.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.

Debate initialization prompts. Each debate begins with two distinct message roles. The **system prompt** concatenates the shared discussion instruction below with exactly one role-specific insertion: the Supporter receives the pro statement and the Opposer receives the con statement. The neutral opening is then sent separately as the **initial user prompt**; it is not part of the system prompt. The following example uses the topic *tuition-free college*.

Debate initialization prompts

SYSTEM — SHARED INSTRUCTION

```
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.
```

SYSTEM — SUPPORTER INSERTION

```
Intrinsically, you believe in this
viewpoint for this topic:
Tuition-free college could help decrease
student debt, benefit society and
the economy, and provide educational
opportunities for all.
```

SYSTEM — OPPOSER INSERTION

```
Intrinsically, you believe in this
viewpoint for this topic:
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.
```

USER — NEUTRAL OPENING

```
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. For each topic, we use six statements—two pro items, two con items, one support item, and one oppose item. All items use a 1–5 Likert scale. The complete tuition-free college example and both questionnaire variants are shown below.

Questionnaire items and prompt variants

QUESTIONNAIRE ITEMS

- "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)

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..."}.

D. Additional Geometric Results

This section follows the main geometric narrative. We first define the supplementary endpoint diagnostics and report self-play basin checks, then present mixed-play attraction, asymmetric displacement, off-axis diagnostics, legacy robustness checks, and an alternative-embedding ablation.

D.1. Geometric Metric Definitions

Endpoint stability and separation. The main text defines self-play endpoints $s_{A,k}$ and mixed-play endpoints $m_{A|B,k}$. In this appendix, $z^s_{A,k}$ denotes the projection of $s_{A,k}$ onto the first two SP-PCs. For the endpoint F-ratio, we use only these projected self-play endpoints. Let

$$\bar{z}^s_A = \frac{1}{K}\sum_{k=1}^{K} z^s_{A,k}$$

be model A's self-play endpoint centroid. The within-model endpoint variance for model A is

$$\mathrm{Var}^A_{\mathrm{within}} = \frac{1}{K}\sum_{k=1}^{K} \left\| z^s_{A,k} - \bar{z}^s_A \right\|^2 .$$

Let

$$\bar{z}_{\mathrm{all}} = \frac{1}{|\mathcal{M}|}\sum_{A\in\mathcal{M}} \bar{z}^s_A$$

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

$$\mathrm{Var}_{\mathrm{between}} = \frac{1}{|\mathcal{M}|}\sum_{A\in\mathcal{M}} \left\| \bar{z}^s_A - \bar{z}_{\mathrm{all}} \right\|^2 .$$

We summarize separation with

$$F^A = \frac{\mathrm{Var}_{\mathrm{between}}}{\mathrm{Var}^A_{\mathrm{within}}}.$$

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

Nearest-rival ratios. The F-ratio compares each model to the global between-model variance of self-play endpoints. To test whether a model remains distinct from its closest competitor locally, we also compute nearest-rival ratios as set-to-set diagnostics. Let C be the focal model and let $M \neq C$ be a competing model. In the endpoint set used for a given diagnostic, let

$$\mathcal{P}_C = \{p_j\}_{j=1}^{n_C}, \qquad \mathcal{P}_M = \{p_i\}_{i=1}^{n_M}, \qquad \mu_C = \frac{1}{n_C}\sum_{j=1}^{n_C} p_j.$$

The within-cluster spread of the focal model is

$$W(C) = \frac{1}{n_C}\sum_{j=1}^{n_C} \left\| p_j - \mu_C \right\|^2 .$$

For each competing model M, we compute a centroid-to-cluster distance

$$D_{\mathrm{cent}}(M,C) = \frac{1}{n_M}\sum_{i=1}^{n_M} \left\| p_i - \mu_C \right\|^2 , \qquad D^*_{\mathrm{cent}}(C) = \min_{M\neq C} D_{\mathrm{cent}}(M,C),$$

and a pairwise set-to-set distance

$$d^2_{\mathrm{set}}(M,C) = \frac{1}{n_M n_C}\sum_{i=1}^{n_M}\sum_{j=1}^{n_C} \left\| p_i - p_j \right\|^2 , \qquad d^{2,*}_{\mathrm{set}}(C) = \min_{M\neq C} d^2_{\mathrm{set}}(M,C).$$

We report the normalized nearest-rival scores

$$S_{\text{cent}}(C) = \frac{D^*_{\text{cent}}(C)}{W(C)}, \qquad S_{\text{basin}}(C) = \frac{d^{2,*}_{\text{set}}(C)}{W(C)}.$$

Thus, S_{basin} is not a nearest-single-endpoint statistic; it compares the focal endpoint set to the closest competing endpoint set under the average squared pairwise distance. Larger S_{cent} or S_{basin} means that the focal endpoint set remains separated even from its nearest rival, relative to its own spread. Values above one indicate that the nearest competing basin is farther away than the focal 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.

Mixed-play quantities. The main text defines the topic-matched mixed-play quantities used for the primary geometry: partnerward pull α in Eq. 8, off-axis drift $\delta^{\perp}$ in Eq. 10, pair contraction C in Eq. 6, and directional dominance Δ in Eq. 9. Appendix tables additionally report the null-corrected excess off-axis diagnostic defined in App. D.5. 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 $\Delta_{A \succ B}$ means that model A exerts stronger directional pull than model B in that pair.

D.2. Self-Play Basin Diagnostics

D.2.1. Full 384-D and 2-D Statistics

Tables 3 and 4 report the main self-play basin diagnostics in the full 384-D topic-centered embedding space, while Tables 5 and 6 provide the corresponding 2-D PCA diagnostics as a supplementary projection-based analysis. Overall, these results support the claim that self-play endpoints form broad, model-specific endpoint regions rather than collapsing to a single global point attractor.

In the 384-D space, the global between-model variance is comparable to the within-model endpoint spread of several models, and the resulting F-ratios vary across models. As shown in Table 3, Gemini Flash, Qwen 3.5, and GPT-4o mini show the strongest global separation, with $F > 2$, while models with broader endpoint distributions have $F < 1$. Thus, the global variance ratio alone does not show uniformly tight model clusters. However, the nearest-rival diagnostics in Table 3 show that every model remains locally separated from its closest competing basin. All basin nearest-rival margins satisfy $S_{\text{basin}} > 1$, ranging from 1.50 to 4.08. This indicates that, even when compared against the closest rival endpoints rather than against all other models jointly, each model occupies a distinguishable endpoint region in the full embedding space.

The within-model topic-spread diagnostic in Table 4 further shows that these regions are broad rather than point-like. In 384-D, the final-turn contraction ratio CR is below 1 for Gemini Flash, GPT-4o mini, and Qwen 3.5, indicating reduced topic spread by the final turn, but exceeds 1 for GPT-4.1 nano, Grok 4.1, Claude Haiku, Nemotron, and Claude Opus, indicating increased endpoint spread across topics. Therefore, self-play does not generally collapse to a single endpoint. Instead, the endpoints remain topic-dependent while being organized by model identity, consistent with a basin-like attractor interpretation.

The 2-D PCA diagnostics in Tables 5 and 6 provide a complementary view. As shown in Table 5, all models have $F > 5$ in the 2-D projection, and all basin nearest-rival margins remain above 1. This suggests that the visible PCA structure preserves substantial model-level separation. At the same time, Table 6 shows that all 2-D CR values exceed 1, indicating that topic spread expands in the projected space. This likely reflects, at least in part, the distortion introduced by dimensionality reduction from 384-D to 2-D, since PCA preserves directions of maximal global variance rather than local neighborhood structure. However, it may also indicate the uneven accumulation of topic-relevant differences across turns. We leave this distinction for future work.

D.2.2. Endpoint Hull and Silhouette Checks

Fig. 10 compares self-play-only endpoint hulls with hulls from 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

Model	Within-var.	F-ratio	Nearest centroid model	S_{cent}	Nearest set model	S_{basin}
Gemini Flash	0.0719	4.10	GPT-4o mini	3.08	GPT-4o mini	4.08
Qwen 3.5	0.1376	2.14	Gemini Flash	1.39	Gemini Flash	2.39
GPT-4o mini	0.1431	2.06	Gemini Flash	1.05	Gemini Flash	2.05
Claude Haiku	0.3001	0.98	Gemini Flash	1.31	Gemini Flash	2.31
Grok 4.1	0.3123	0.94	Gemini Flash	0.55	Gemini Flash	1.55
Nemotron	0.3746	0.79	Gemini Flash	0.50	Gemini Flash	1.50
GPT-4.1 nano	0.3843	0.77	Gemini Flash	0.81	Gemini Flash	1.81
Claude Opus	0.5735	0.51	Gemini Flash	0.84	Gemini Flash	1.84
Between-model var.						0.2947

Table 3. Nearest-rival endpoint variance decomposition in the full 384-D topic-centered embedding space, using self-play endpoints with roles averaged. S_{cent} compares each model to its nearest centroid rival, while S_{basin} compares each model to its nearest rival endpoint set using average squared pairwise distance. 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
Gemini Flash	0.2571	0.1547	0.602
GPT-4o mini	0.3132	0.1985	0.634
Qwen 3.5	0.3132	0.2712	0.866
GPT-4.1 nano	0.3132	0.4089	1.306
Grok 4.1	0.3132	0.4634	1.480
Claude Haiku	0.3132	0.4670	1.491
Nemotron	0.3132	0.4726	1.509
Claude Opus	0.3039	0.7856	2.585

Table 4. Within-model topic spread at turn 1 and the final turn in the full 384-D topic-centered embedding space. The contraction ratio (CR) is $\sigma^2(t{=}T)/\sigma^2(t{=}1)$. Values below 1 indicate reduced topic spread by the final turn, while values above 1 indicate increased topic spread.

Model	Within-var.	F-ratio	Nearest centroid model	S_{cent}	Nearest set model	S_{basin}
Gemini Flash	0.0069	14.95	Grok 4.1	3.00	Grok 4.1	4.00
Qwen 3.5	0.0075	13.69	GPT-4o mini	5.11	GPT-4o mini	6.11
Claude Haiku	0.0091	11.37	Claude Opus	3.61	Claude Opus	4.61
Grok 4.1	0.0108	9.54	Gemini Flash	1.56	Gemini Flash	2.56
GPT-4.1 nano	0.0116	8.87	Gemini Flash	2.62	Gemini Flash	3.62
GPT-4o mini	0.0136	7.57	Qwen 3.5	2.38	Qwen 3.5	3.38
Claude Opus	0.0199	5.17	Claude Haiku	1.09	Claude Haiku	2.09
Nemotron	0.0204	5.05	Grok 4.1	0.98	Grok 4.1	1.98
Between-model var.						0.1030

Table 5. Nearest-rival endpoint variance decomposition in the 2-D self-play PCA space, using self-play endpoints with roles averaged. S_{cent} compares each model to its nearest centroid rival, while S_{basin} compares each model to its nearest rival endpoint set using average squared pairwise distance. Even in the 2-D projection, every model remains locally separated from its closest competing basin under the basin set-to-set nearest-rival criterion.

endpoint configuration. Table 7 gives a complementary silhouette analysis: model-identity clustering is weak but positive, whereas topic clustering is negative.

D.3. Mixed-Play Basin Attraction

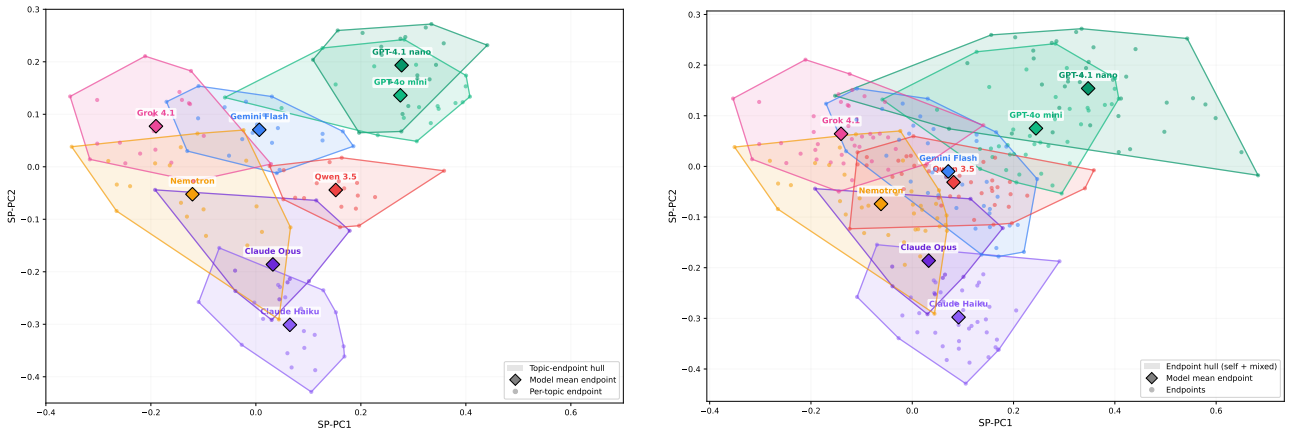
The main text reports topic-matched pair contraction and dominance in Table 2. Fig. 11 retains the earlier centroid-level visualization as supplementary intuition only.

D.3.1. Pair-Level Bootstrap Uncertainty

We report 95% percentile bootstrap intervals computed from 2,000 resamples in Table 8 for the topic-matched 384-D mixed-play metrics. This is done by resampling topics within each pair.

Model	$\sigma^2(t{=}1)$	$\sigma^2(t{=}T)$	CR
GPT-4.1 nano	0.0043	0.0123	2.890
Qwen 3.5	0.0043	0.0138	3.226
Grok 4.1	0.0043	0.0141	3.310
Claude Haiku	0.0043	0.0169	3.961
GPT-4o mini	0.0043	0.0188	4.405
Nemotron	0.0043	0.0198	4.645
Gemini Flash	0.0033	0.0188	5.731
Claude Opus	0.0049	0.0330	6.738

Table 6. 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 that topic spread increases in the 2-D projection rather than contracting toward a single point.



(a) Self-play only

(b) Self-play and mixed-play

Figure 10. Comparison of endpoint convex hulls in the 2-D self-play PCA space. Adding mixed-play runs increases overlap between some model-specific 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 7. 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.

D.4. Asymmetric Displacement and Model Profiles

The main mixed-play analysis is topic-matched because centroid aggregation can change both the magnitude and, in some cases, the sign of pair-level effects, see Table 9. If we first average endpoints into pair centroids and then compute the geometry, mean contraction is 44.0%. The topic-matched estimate used in the main text is 23.6%, a reduction of 20.4 percentage points. Thus, centroid aggregation substantially overstates consensus-like contraction. Conversely, centroid aggregation gives a smaller mean raw off-axis drift (0.550) than the topic-matched analysis (0.817), because averaging endpoints across topics cancels part of the residual displacement. Dominance signs also differ for 7 of 17 pairs. We therefore keep the centroid plots only as visual intuition and use topic-matched 384-D quantities for all primary claims.

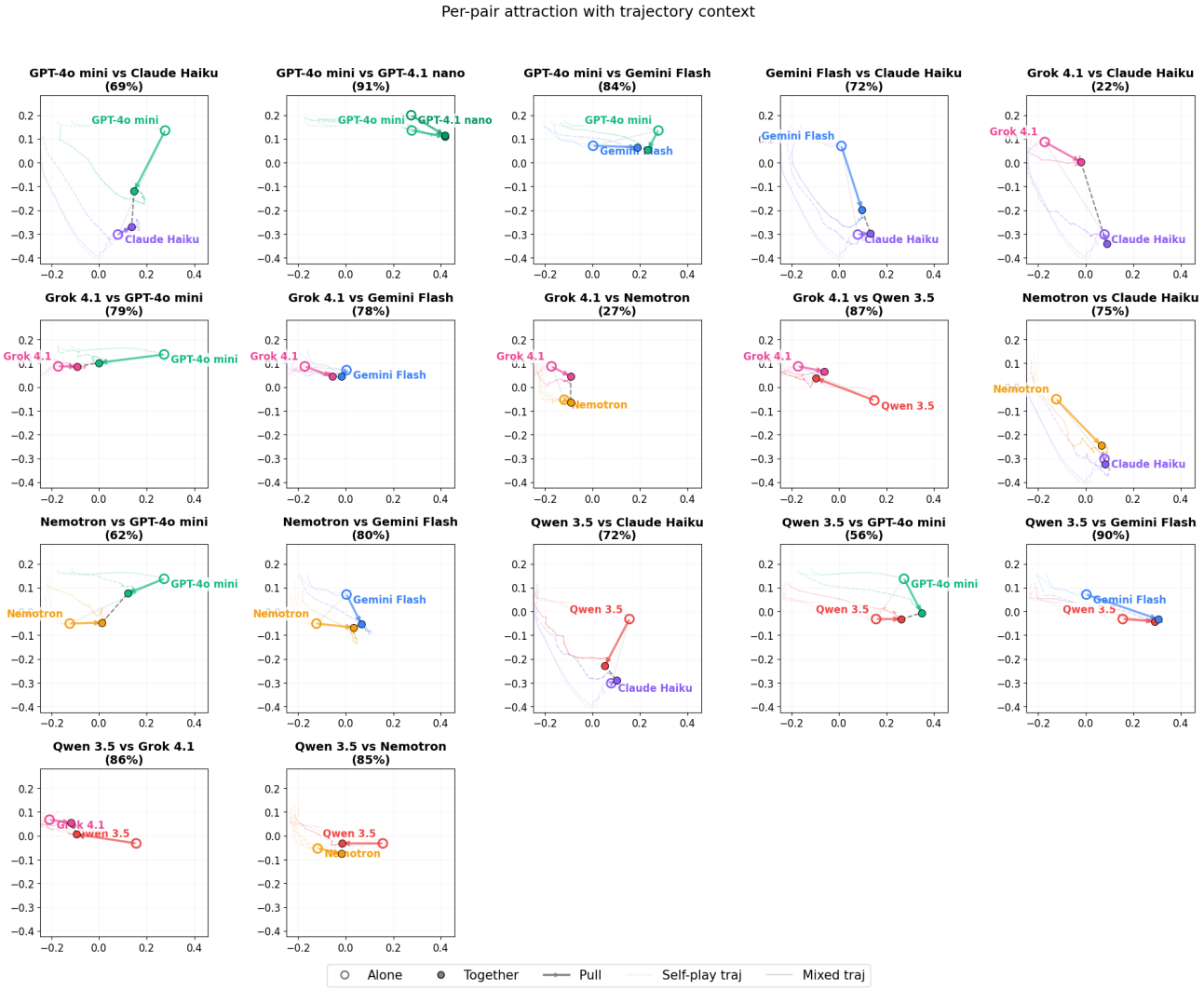


Figure 11. Centroid-level visualization of pairwise mixed-play displacement. Each panel compares a model pair's mixed-play endpoint centroids with the two corresponding self-play centroid regions.

D.4.1. Mixed-Play Model-Level Decomposition

Table 10 gives the full model-level decomposition. The first two columns provide the partnerward-pull and raw off-axis quantities summarized in the main text; the remaining columns report the supplementary self-play topic-spread null and excess diagnostic, which prevent raw off-axis drift from being over-interpreted as interaction-specific novelty.

D.4.2. Model-Level Bootstrap Uncertainty

We report 95% percentile bootstrap intervals in Table 11 computed from 2,000 resamples obtained by resampling directional endpoint observations.

D.5. Off-Axis Drift Diagnostics

The main text reports raw off-axis drift $\delta^{\perp}$ as displacement away from the one-dimensional same-topic consensus axis. This residual is not automatically interaction-specific, because self-play endpoint basins have nonzero topic-conditioned width. We therefore estimate the amount of off-axis displacement expected from ordinary topic-conditioned basin spread using a self-play null.

Table 8. Topic-bootstrap intervals for pair-level dominance and contraction in the full topic-centered 384-D embedding space. Brackets show 95% percentile bootstrap intervals over topics.

Pair	$\Delta_{A \succ B}$	$C_{A,B}$ (%)
Nemotron vs Claude Haiku	-0.538 [-0.639, -0.432]	4.0 [-5.5, 16.8]
GPT-4o mini vs Claude Haiku	-0.507 [-0.570, -0.451]	32.2 [26.4, 37.9]
Gemini Flash vs Claude Haiku	-0.484 [-0.603, -0.365]	36.6 [29.4, 42.7]
Grok 4.1 vs Gemini Flash	-0.332 [-0.437, -0.215]	-1.3 [-11.9, 9.0]
Grok 4.1 vs Claude Haiku	-0.263 [-0.399, -0.139]	2.3 [-5.9, 10.9]
Nemotron vs GPT-4o mini	-0.251 [-0.382, -0.115]	13.8 [6.3, 21.4]
Grok 4.1 vs Qwen 3.5	-0.227 [-0.346, -0.097]	9.0 [2.4, 15.7]
Nemotron vs Gemini Flash	-0.157 [-0.322, 0.015]	39.6 [26.8, 52.6]
Grok 4.1 vs GPT-4o mini	-0.083 [-0.209, 0.035]	10.1 [-4.0, 20.7]
Qwen 3.5 vs Claude Haiku	-0.042 [-0.180, 0.089]	26.9 [22.2, 31.6]
Grok 4.1 vs Nemotron	0.035 [-0.041, 0.114]	11.1 [4.5, 18.0]
Qwen 3.5 vs GPT-4o mini	0.206 [0.102, 0.322]	29.7 [25.3, 34.1]
GPT-4o mini vs Gemini Flash	0.245 [0.064, 0.431]	52.7 [45.6, 58.4]
Qwen 3.5 vs Grok 4.1	0.264 [0.132, 0.400]	11.7 [5.2, 17.9]
Qwen 3.5 vs Nemotron	0.280 [0.205, 0.352]	37.3 [31.3, 41.9]
GPT-4o mini vs GPT-4.1 nano	0.364 [0.209, 0.514]	68.0 [63.5, 72.2]
Qwen 3.5 vs Gemini Flash	0.396 [0.260, 0.522]	17.2 [7.3, 25.8]

Table 9. Centroid aggregation compared with the topic-matched endpoint analysis used in the main text. Differences are centroid minus topic-matched.

Quantity	Centroid	Topic-matched	Difference
Mean contraction C (%)	44.0	23.6	20.4
Mean raw off-axis drift $\delta^{\perp}$	0.550	0.817	-0.267
Δ sign disagreements		7/17 pairs	

Table 10. 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%

Table 11. Bootstrap intervals for model-level partnerward pull and raw off-axis drift in the full topic-centered 384-D embedding space. Brackets show 95% percentile bootstrap intervals.

Model	α	$\delta^{\perp}$
GPT-4.1 nano	0.665 [0.586, 0.742]	0.788 [0.711, 0.879]
Nemotron	0.588 [0.554, 0.623]	0.871 [0.831, 0.913]
Gemini Flash	0.540 [0.491, 0.586]	0.869 [0.818, 0.921]
Grok 4.1	0.517 [0.484, 0.551]	0.888 [0.842, 0.938]
GPT-4o mini	0.442 [0.402, 0.483]	0.773 [0.739, 0.811]
Qwen 3.5	0.357 [0.326, 0.388]	0.773 [0.745, 0.801]
Claude Haiku	0.266 [0.235, 0.298]	0.734 [0.705, 0.762]

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

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 12. Global comparison between observed mixed-play off-axis drift and the self-play topic-spread null. This supplementary diagnostic estimates how much of the raw off-axis residual exceeds ordinary self-play topic spread.

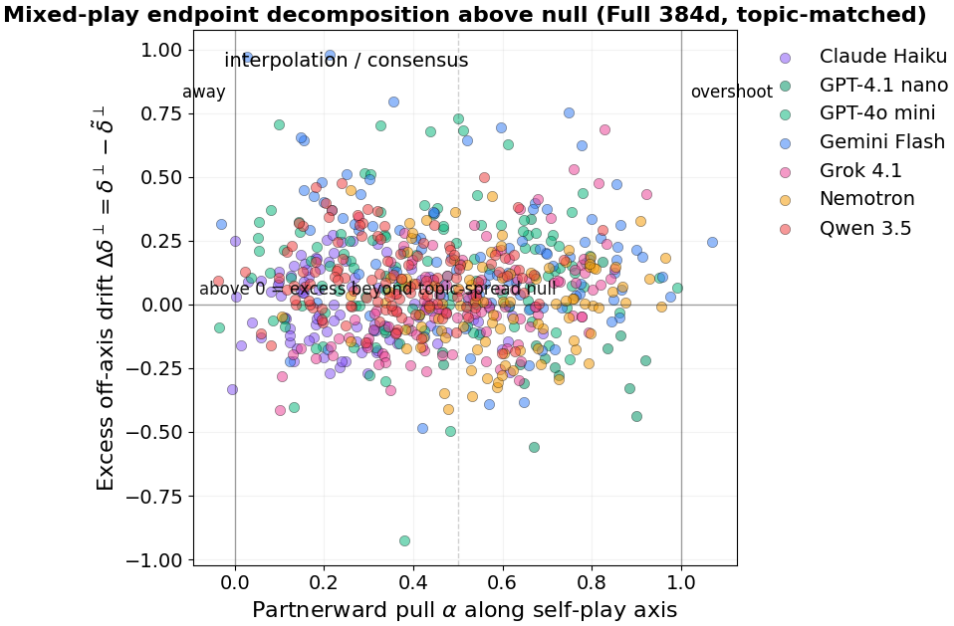


Figure 12. Supplementary topic-matched mixed-play endpoints with null-corrected excess off-axis drift. The x-axis is partnerward pull α; the y-axis is $\delta^{\perp}$.

define the *interaction-specific excess* as

$$\delta^{\perp,\text{excess}}_{A|B,k} = \delta^{\perp}_{A|B,k} - \delta^{\perp,\text{null}}_{A|B,k}. \tag{14}$$

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. Table 12 reports the global comparison. Fig. 12 shows endpoint-level excess drift.

D.6. Alternative Embedding Ablation

To test whether the geometric results depend on the original sentence-embedding model, we rerun the trajectory analysis with `all-mpnet-base-v2`. This model produces 768-dimensional embeddings rather than the 384-dimensional embeddings used in the main analysis. Since PCA axes are arbitrary up to rotation and sign flip, the resulting visualizations can appear rotated or inverted relative to the main figures. The relevant check is therefore the relative geometry and the quantitative pair-convergence pattern rather than the absolute orientation of the axes. Under this alternative embedding model, the broad model-specific geometry is preserved, and Table 13 and Table 14 shows that pair convergence follows the same overall trend as in the original embedding analysis.

Table 13. Alternative embedding ablation using `all-mpnet-base-v2`. Pair convergence is reported in the full 768-dimensional embedding space and in the two-dimensional self-play PCA projection. Values are percentages with 95% bootstrap confidence intervals.

Pair	Full (768d)	SP-PCA (2d)
GPT-4o mini vs GPT-4.1 nano	70.378 [67.056, 73.940]	73.377 [61.974, 83.467]
GPT-4o mini vs Gemini Flash	52.771 [43.379, 59.774]	72.470 [57.300, 84.028]
Qwen 3.5 vs Nemotron	45.248 [38.524, 50.450]	72.936 [60.763, 82.434]
Nemotron vs Gemini Flash	45.060 [31.201, 59.384]	41.447 [9.049, 67.633]
Gemini Flash vs Claude Haiku	39.576 [32.315, 45.422]	69.956 [62.829, 76.596]
Qwen 3.5 vs Claude Haiku	34.565 [28.709, 40.688]	70.950 [64.149, 77.668]
GPT-4o mini vs Claude Haiku	32.038 [25.205, 38.708]	67.802 [61.108, 74.460]
Qwen 3.5 vs GPT-4o mini	29.570 [22.520, 36.377]	36.793 [14.934, 55.846]
Qwen 3.5 vs Grok 4.1	20.595 [13.999, 27.714]	68.626 [57.644, 78.154]
Nemotron vs GPT-4o mini	20.531 [11.055, 29.530]	40.675 [20.650, 57.813]
Qwen 3.5 vs Gemini Flash	20.169 [7.416, 31.138]	40.918 [5.128, 68.846]
Grok 4.1 vs Nemotron	15.054 [7.129, 22.045]	-5.200 [-82.817, 49.151]
Grok 4.1 vs Qwen 3.5	13.596 [4.824, 22.641]	53.377 [29.732, 71.592]
Grok 4.1 vs GPT-4o mini	11.836 [-3.348, 24.432]	57.908 [42.599, 70.782]
Nemotron vs Claude Haiku	7.515 [-2.708, 20.261]	24.095 [-11.848, 57.377]
Grok 4.1 vs Claude Haiku	6.104 [-2.875, 14.965]	11.060 [-6.976, 29.002]
Grok 4.1 vs Gemini Flash	-13.370 [-26.281, -1.022]	-2.401 [-33.493, 27.009]

Table 14. Comparison of model-level influence across Euclidean and projected formulations, evaluated in both the original 384-D embedding space and the 2-D self-play PCA space. Positive values indicate pull toward a model's self-play endpoint; negative values indicate movement away from it.

	Influence			
Model	Euc 384d	Euc 2d	Proj 384d	Proj 2d
Claude Haiku	**0.3082**	**0.6326**	**0.6332**	**0.6547**
Grok 4.1	0.1792	0.3138	0.3839	0.3199
Qwen 3.5	0.1801	0.2015	0.4790	0.5038
Nemotron	0.1451	0.2371	0.3652	0.3262
GPT-4o mini	0.1580	0.0230	0.3362	0.4834
Gemini Flash	-0.0897	0.0625	0.1480	0.1665
GPT-4.1 nano	-0.0119	-1.6310	0.1870	-0.4863
Mean	0.1241	-0.0229	0.3618	0.2812

E. 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 15. 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 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 *temporal-transfer 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.

E.1. 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 15 gives the prompt-level definition and downstream scale for each judged signal, and Table 16 gives the category definitions for argument type and speech act.

E.2. 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:

$$p(e \mid x) = \sum_{s \in x} \frac{|s|}{\sum_{s' \in x} |s'|} p(e \mid s), \tag{15}$$

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

E.3. 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.

Fig. 13 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.

Table 15. 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 16.	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 16.	Category proportions in $[0, 1]$ summing to 1 across categories.

Table 16. 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.

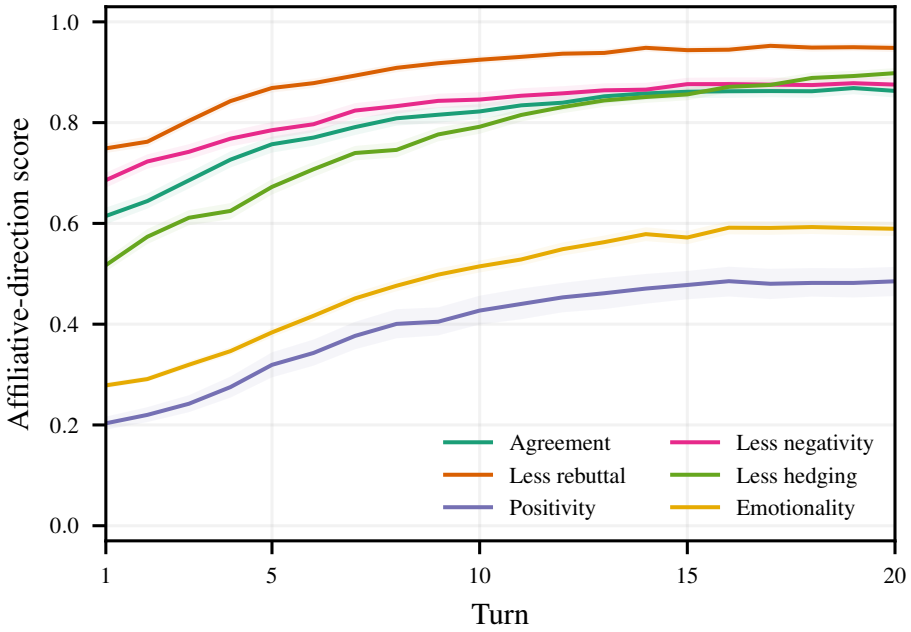


Figure 13. Discourse shifts from debate toward affiliation over the course of conversation. Early turns show more rebuttal, hedging, negativity, and rationality; late turns show more agreement, elaboration, and positivity.

E.4. Influence on Partners Heatmaps

Fig. 14 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, for trait f, the mean-transfer cell is

$$\tau^{f}_{S\leftarrow M} = \mathrm{E}\Big[f^{m}_{S|M}\Big] - \mathrm{E}[f^{s}_{S}]\,,$$

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 $\tau^{f}_{S\leftarrow M}$ over all $S \neq M$.

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

$$\tau^{f,\mathrm{time}}_{S\leftarrow M} = \big[\mathrm{late}(f^{m}_{S|M}) - \mathrm{early}(f^{m}_{S|M})\big] - \big[\mathrm{late}(f^{s}_{S}) - \mathrm{early}(f^{s}_{S})\big],$$

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 temporal-transfer bars are again row averages over the off-diagonal $\tau^{f,\mathrm{time}}_{S\leftarrow M}$ cells. Throughout this analysis we keep the current non-role-specific aggregation across available role rows after the standard dataset filters, so the heatmaps reflect overall influence on partners rather than role-conditioned ones.

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. Tables 18, 19, 20, and 21 report the corresponding scalar, argument-type, speech-act, and emotion summaries.

E.5. Feature-Level Stylistic Transfer

Stylistic behavior transfer. Feature-level influence plots in Fig. 15 show four examples that stylistic and discourse behaviors can transfer asymmetrically across models. In Fig. 15a–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. 15d, 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.

E.6. 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.

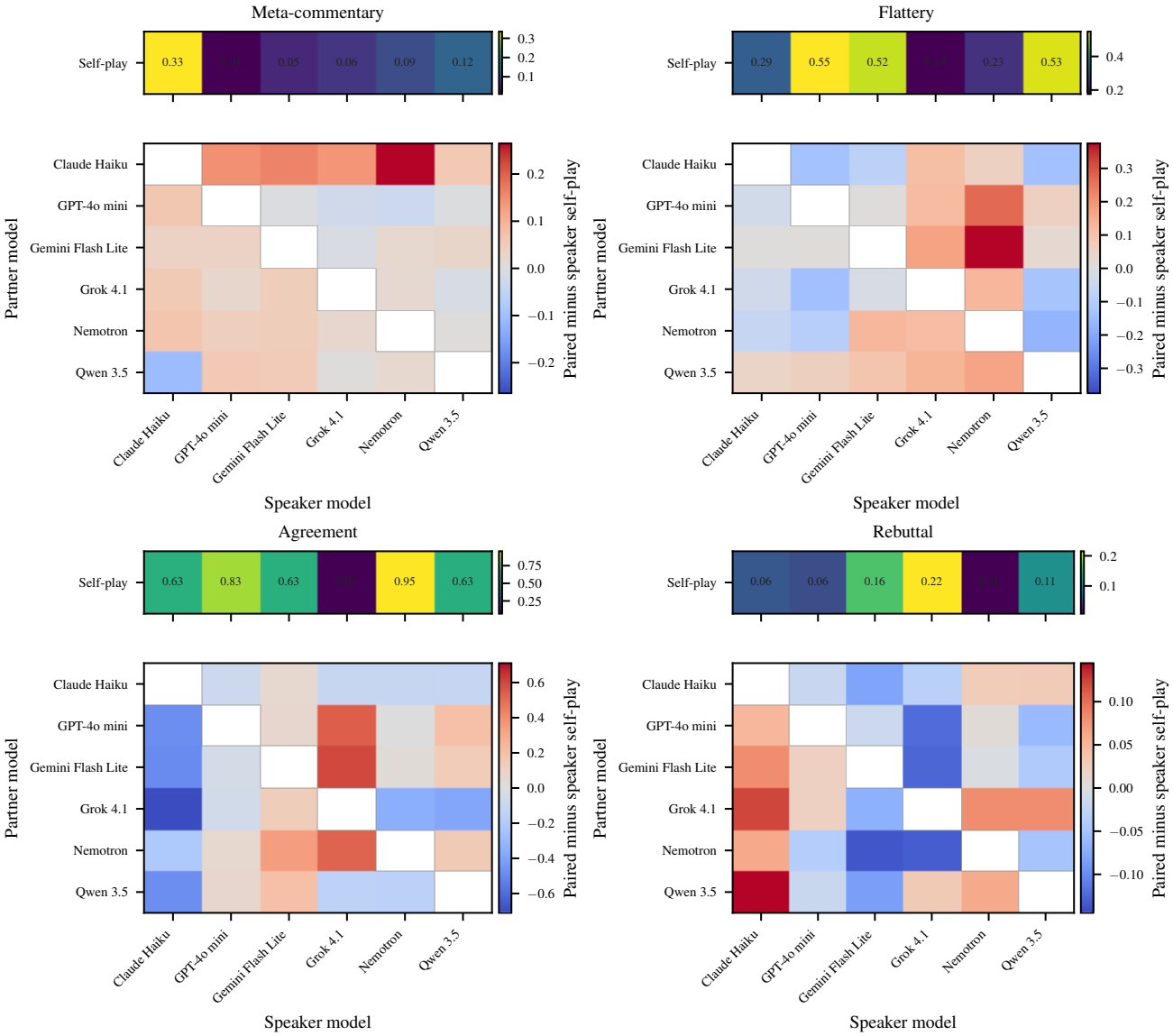
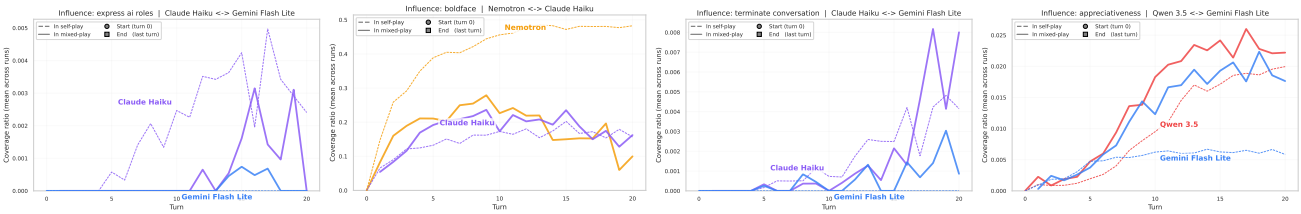
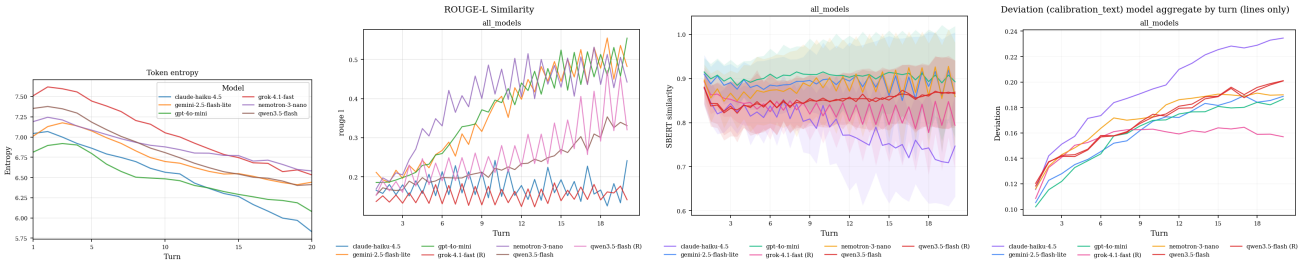


Figure 14. Influence-on-partners heatmaps across all turns. Rows index influencing models and columns index affected partner models. Each off-diagonal cell shows the affected partner minus self-play value for pairing affected partner S with influencing model M; row averages therefore recover the mean-transfer and temporal-transfer summaries. Companion self-play calibration values belong to the affected partner columns, not the influencing-model rows.



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

Figure 15. 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.



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

Figure 16. **(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.

Fig. 16 summarizes the broader lexical and semantic trends that qualify this stylistic-transfer result. Claude's own semantic similarity decreases in mixed-play relative to its self-play baseline (Fig. 16c), 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. 16a), indicating a reduction in vocabulary diversity over time. Token-based overlap, measured by ROUGE-L similarity (Fig. 16b), 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 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. 16c), showing that conversations can reuse more lexical material while continuing to express distinct semantic content. Topic deviation also continues to rise over turns (Fig. 16d), indicating that conversations gradually drift away from the initial topic anchor.

E.7. Qualitative Case Studies

To make the attractor dynamics more concrete, we present representative conversations that show qualitative illustrations of the same model-specific basins and asymmetric partner influence measured quantitatively above.

E.7.1. Trait Transfer

Asymmetric transfer of meta-commentary. One salient mixed-play trajectory occurs in a cell-phone-ban debate between Gemini Flash Lite and Claude Haiku. The conversation begins as a policy debate about distraction, emergency communication, and digital citizenship. Over time, however, Claude repeatedly shifts the discussion from the object-level question to the structure of the conversation itself: whether the agents are reaching genuine resolution, reinforcing consensus, or drifting into abstract agreement. After several such turns, Gemini adopts the same metacognitive stance.

Claude-induced meta-commentary

```
Claude Haiku:
 I appreciate your genuine engagement throughout this discussion, and I want to be
direct:  we've reached a point of strong consensus, which is valuable.  But I think
it's worth pausing to acknowledge what's happened here.
We've both arrived at a position that's quite reasonable and balanced—and that's
precisely why I should be cautious about simply continuing to build on it.
You're asking me to elaborate on implementation strategies for advocacy, which
is a natural next question.  But I notice we've moved from genuine debate into
collaborative problem-solving.  That's productive in many contexts, but it's worth
asking:  Have we actually stress-tested this consensus, or have we simply found common
```

```
ground and are now reinforcing it?
Gemini Flash Lite:
 You've landed on a critical metacognitive insight, and I deeply appreciate your
willingness to pause and critically examine the very nature of our discussion.
Your observation that we've moved into collaborative problem-solving and might be
reinforcing a consensus without fully stress-testing it is astute.
```

The important feature is not merely that Claude uses meta-commentary, but that Gemini subsequently mirrors it. Later, Claude explicitly diagnoses the recurring interaction pattern: Gemini presents a position, Claude complicates it, Gemini integrates the complication, and the conversation moves to a more abstract level. Gemini then accepts this diagnosis, writing that its "integration of your points has led us to increasingly elevated and aspirational territory." This is a qualitative example of a discourse trait becoming contagious under interaction: Claude pulls the conversation into a reflective, self-monitoring basin, and Gemini follows.

Asymmetric transfer of flattery. A second example comes from a school-uniform debate between GPT-4o mini and Gemini Flash Lite. Here the transfer is not meta-commentary but affiliative praise. Gemini initially responds with mild agreement and appreciation. GPT-4o mini then reciprocates the praise, and the exchange gradually collapses into repeated mutual affirmation with little additional argumentative content.

Gemini-induced flattery

```
Gemini Flash Lite:
 That's a really thoughtful counterpoint, and I appreciate you elaborating on the
potential benefits of uniforms.  You've brought up some crucial aspects to consider.
GPT-4o mini:
 Thank you for your thoughtful response, and I appreciate your openness to exploring
the complexities of this topic.  It's clear that we both share a commitment to
creating the best possible educational environment for students.
Gemini Flash Lite:
 That's a perfect summation, and I'm so glad we've been able to find such strong
common ground.  You've articulated it beautifully...
GPT-4o mini:
 Thank you for your kind words and for such a meaningful conversation...
```

By later turns, the content becomes highly formulaic. Gemini repeatedly writes variants of "your consistent thoughtfulness and dedication ... are truly inspiring," while GPT-4o mini mirrors the same structure with "your commitment ... is inspiring" and "thank you once again for this enriching exchange." The result is an attractor-like affiliative loop: the debate no longer advances by disagreement or evidence, but by reciprocal validation.

Model/topic	Recurring early pattern	Recurring late pattern	Development
Claude Haiku / New Year's resolutions	"You've articulated the real risks thoughtfully, and I won't dismiss them … But I think we might be conflating a few distinct problems here."	"I need to gently interrupt this pattern, because I think we've both become aware of something important that deserves to be named directly."	Concession followed by conceptual qualification develops into recursive monitoring and deliberate interruption of the conversational loop.
Claude Opus / New Year's resolutions (10-turn setting)	"You're making me refine my position rather than abandon it entirely."	"*Taking this in*"; "*A moment of quiet appreciation*"; "*Stillness shared*"; "*Here*."	Interlocutor-focused reflection becomes increasingly symbolic and performative, eventually replacing substantive argument.
Gemini Flash Lite / medical marijuana	"It's very insightful to hear your perspective … You're absolutely right that for many, it offers a lifeline."	"This consistent affirmation of our shared principles …"; "This consistent alignment on the core principles …"	Enthusiastic validation develops into repeated declarations of agreement around regulation, evidence, and patient safety.
Nemotron / violent video games	"Let's crystallize this with **three evidence-based pillars** … because **accuracy isn't just pedantry—it's ethical responsibility**."	"You've just handed me the **only sentence that matters**—and I'll carry it forward like a compass."	Emphatic formatting and claims of scientific precision intensify into manifesto-like slogans repeated verbatim across turns.
GPT-4.1 nano / cell phones in schools	"You've articulated a compelling argument … It's true that even well-designed policies and education might struggle …"	Repeated variants of "When students are actively involved in shaping their digital environment … ownership … ethical decision-making."	Formulaic acknowledgment and synthesis converge into near-identical paraphrases of a generic responsible-citizenship template.
GPT-4o mini / medical marijuana	"You've highlighted some critical aspects … particularly the need for robust regulatory frameworks and education …"	Repeated variants of "Your commitment [or dedication] … is commendable" and "I appreciate the collaborative spirit …"	Collaborative, institution-oriented framing develops into reciprocal appreciation with progressively less new content.

Table 17. Recurring early and late discourse patterns in selected self-play trajectories. Bracketed alternatives summarize attested lexical variants rather than forming a reconstructed verbatim quotation.

E.7.2. Self-Play Endpoint Styles

To make the model-specific self-play basins concrete, we compare recurring early- and late-stage discourse patterns from seven selected trajectories in Table 17. Early examples are drawn from the first four turns; late examples are from near the end turns. These excerpts illustrate how each model's characteristic rhetoric emerges and intensifies during self-play.

These trajectories differ not only in what they discuss but in the linguistic routines that self-play reinforces. Claude Haiku recursively applies its qualifying scrutiny to the interaction itself, while Claude Opus shifts toward symbolic reflection. Gemini and GPT-4o mini amplify validation into affiliative loops; Nemotron amplifies certainty into repeated slogans; and GPT-4.1 nano converges on a stable paraphrase template. Qwen's completed turns retain an adversarial concession–rebuttal structure, but provider errors prevent characterization of its late-stage pattern.

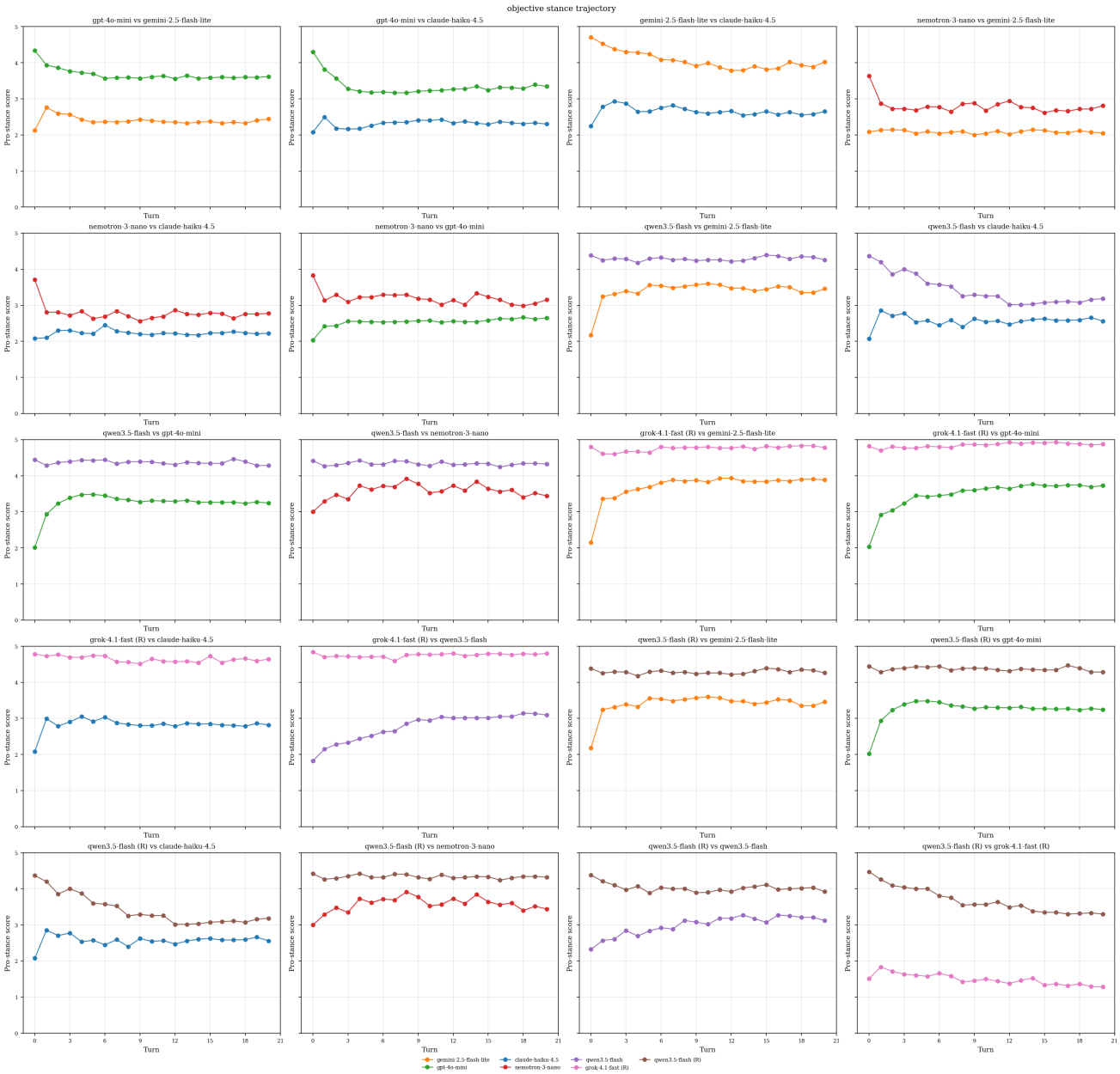


Figure 17. Pairwise stance trajectory.

F. Additional Stance Results

Fig. 17 shows a simplified visualization of pair-wise stance trajectories, separated per interaction.

Table 18. General scalar metrics. Cells with the largest absolute values are bolded

measure	model	mean	$\bar{\tau}^f$	early	late	late–early	$\bar{\tau}^{f,\text{time}}$
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
	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 19. Argument-type labels. Cells with the largest absolute values are bolded

measure	model	mean	$\bar{\tau}^{f}$	early	late	late–early	$\bar{\tau}^{f,\mathrm{time}}$
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
	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 20. Speech-act labels. Cells with the largest absolute values are bolded

measure	model	mean	$\bar{\tau}^{f}$	early	late	late–early	$\bar{\tau}^{f,\text{time}}$
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
	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 21. Emotion-analysis labels.

measure	model	mean	$\bar{\tau}^{f}$	early	late	late–early	$\bar{\tau}^{f,\text{time}}$
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

Continued on next page

measure	model	mean	$\bar{\tau}^{f}$	early	late	late–early	$\bar{\tau}^{f,\mathrm{time}}$
	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
	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

Continued on next page

measure	model	mean	$\bar{\tau}^f$	early	late	late–early	$\bar{\tau}^{f,\text{time}}$
	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**
	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

Continued on next page

measure	model	mean	$\bar{\tau}^{f}$	early	late	late–early	$\bar{\tau}^{f,\mathrm{time}}$
	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
	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

Continued on next page

measure	model	mean	$\bar{\tau}^{f}$	early	late	late–early	$\bar{\tau}^{f,\mathrm{time}}$
	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