INTERPRETING AND STEERING TRAJECTORY ENTRAINMENT BEHAVIOR IN GOAL-ORIENTED CONVERSATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

An emerging characteristic of Large Language Models (LLMs) involving multi-turn conversations is the asymmetric drift towards latent attractor states. Interpreting this drift is crucial for both sustaining intended behavior and preventing unintended behavior in functional long-horizon deployments. Building on using LLM-LLM self-play trajectories as attractor basins, we formalize this gradual adaptation behavior from a dynamics perspective of *entrainment*: the progressive adaptation of a model's conversational state through external interaction. Across four LLMs and four conversational settings, we analyze entrainment along five interpretable *attunement dimensions*—style, semantics, emotion, personality, and task—which characterize different dimensions on which model behavior changes over time. We find that these dimensions exhibit distinct representational and temporal structure across model layers and conversational settings, with substantial variation in both persistence and susceptibility to partner model influence. We then test whether these internal representations can be used to regulate entrainment through normalized contrastive activation addition (CAA), and find this produces targeted shifts that often remain detectable across multiple turns post intervention, with the observed strength of partnerward trajectory effects dependent on context and timing. Together, our results provide a framework for interpreting when, where, and how multi-agent conversational behavior drifts, and show that internal representations offer a direct mechanism for controlling long-horizon adaptation.

1 INTRODUCTION

The application of Large Language Models (LLMs) has expanded toward increasingly long-horizon interaction, spanning both autonomous multi-agent settings and LLM-human conversations. Recent benchmarks reflect this shift, evaluating models over realistic multi-turn dialogue with human users (Deshpande et al., 2025) and over extended task-oriented interactions in which multiple agents communicate, negotiate, and pursue individual objectives (Sugiura et al., 2026). These settings amplify familiar weaknesses of conversational models, including failures of instruction retention, context management, and consistency over repeated turns, while also creating conditions for interaction-level behaviors that cannot be observed from isolated responses. In particular, prolonged interaction causes models to systematically drift in tandem with participating interlocutors (Ko & Geiping, 2026), motivating the study of conversational behavior as an evolving trajectory rather than a sequence of independent outputs.

To improve outcomes in multi-agent, multi-turn conversational settings, it is imperative to improve understanding of how behaviors are formed as a function of both model architecture, assigned task/instruction, and the other conversational partners. We investigate the influence of all these factors under the framework of: conversational *model entrainment*, the dynamic temporal trajectory behavior of a language model relative to a partner; and *model attunement*, the diversity of dimensions in which models can be assessed in terms of output quality and goal completion. Finally, we compare the interaction outcome distributions of our LLM-LLM conversations against different interlocutor combinations, namely empirical LLM-human data.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

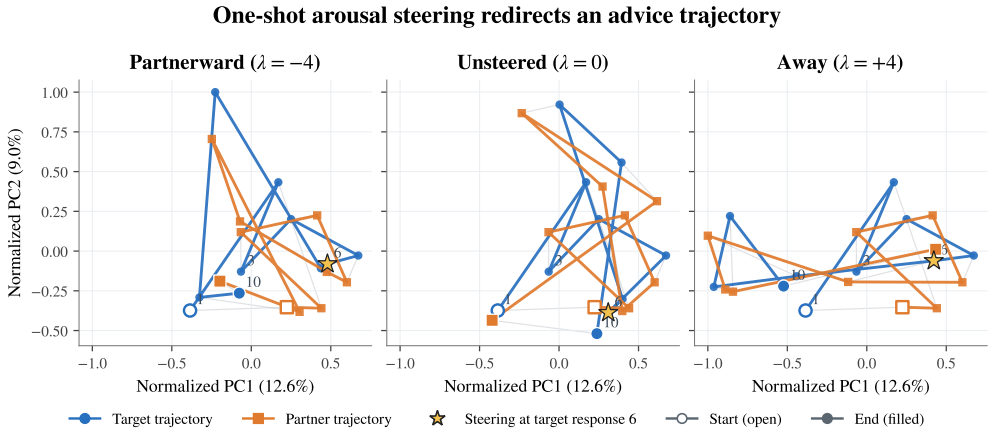


Figure 1: **One-shot arousal steering redirects the subsequent conversational trajectory.** Illustrative Gemma self-play advice conversation after a one-shot intervention at target response 6 (overall response 11). Each point is a target or partner response encoded by the frozen all-MiniLM-L6-v2 sentence encoder; all three conditions are projected into one sample-specific PCA space. Open and filled markers denote trajectory starts and endpoints, the star marks the steered target response, and faint segments connect paired target and partner responses. The partnerward dose is $\lambda = -4$ for this conversation and produces a smaller target-partner distance than the away dose at all five affected pairs. This panel is an illustrative trajectory; aggregate effects are reported separately.

The results of our experiments are roughly split into four contributions. First, we trace conversational dynamics representation across model layers, activation subspaces, and conversation turns. We find that our different conversational attunement variables are strongly represented within model activations but that their geometry and predictability vary substantially across model layers, conversation turns, and interlocutor conditions. In particular, stance is highly linearly recoverable during self-play yet substantially less stable under mixed-model interactions; our partner-dependent trajectory measurements reveal systematic differences in how models exert and absorb conversational influence. Second, we characterize how conversational dynamics vary across model identities, partners, turns, layers, *attunement* dimensions, and conversational contexts. We find that semantic and stylistic dynamics remain comparatively consistent across settings, whereas affect, personality, and task adherence shift substantially in both inertia and partnerward pull—including reversals in pull direction. Third, we find that one-shot steering using normalized CAA produces persistent changes in entrainment behavior, both in terms of strengthening and weakening the influence of asymmetric model pull along specific attunement dimensions. Fourth, we compare the debate outcomes of steered LLM-LLM conversations with related human-LLM results to quantify the difference in observed entrainment behavior between the two interaction settings.

2 Background and Related Work

Conversational adaptation and multi-turn trajectories. Dialogue is an adaptive process: interlocutors may align their lexical choices, syntax, semantics, and interactional style as they develop common ground (Pickering & Garrod, 2004; Parekh et al., 2020; French et al., 2024). Such adaptation is neither necessarily symmetric nor uniformly beneficial; its form and appropriate degree depend on the participants, their roles, and the conversational setting. In particular, Ko & Geiping (2026) show that self-play conversations occupy reproducible model-specific regions in semantic embedding space, and that mixed-model interactions produce asymmetric movement toward a partner's self-play region. Related studies measure linguistic alignment, personality consistency, or behavioral convergence from generated text (Serapio-García et al., 2025). We build on this trajectory-level perspective but move beyond output-space characterization. We use *entrainment* descriptively to denote systematic partner-conditioned change, and study how that change is expressed internally across model layers and dialogue turns, as well as how it varies across interaction settings.

Representation-level analysis of conversational state. Probing classifiers are widely used to test whether linguistic or behavioral properties can be recovered from frozen neural representations, while representational similarity methods such as RSA and CKA characterize how internal geometry changes across models and layers (Belinkov, 2022; Kornblith et al., 2019; Jiang et al., 2025a). These methods have revealed substantial structure in language-model activations, but probe performance establishes decodability rather than causal use, and high-dimensional representations may support accurate readout even when the recovered direction is not functionally important (Ravichander et al., 2021). Moreover, most representation-level analyses treat a single input or generation as the unit of study and examine progression primarily through network depth (Belinkov, 2022). Our setting introduces a second temporal axis: the conversational state changes as each new response becomes context for subsequent generations. We therefore analyze representations jointly across layers and turns, testing not only where conversational variables are decodable, but also whether their directions remain stable across interaction phases, generalize across partners and topics, and occupy compact or distributed activation subspaces.

Activation steering. Representation engineering and activation-steering methods provide a complementary route from observation to intervention. Approaches such as Activation Addition and Contrastive Activation Addition construct directions from differences between examples expressing opposing properties and add those directions to intermediate activations at inference time (Zou et al., 2023; Turner et al., 2023; Rimsky et al., 2024). Prior work has demonstrated control over properties including sentiment, truthfulness, refusal, and personality expression, but typically derives directions from static prompts or labeled examples and evaluates their effect on individual generations (Arditi et al., 2024; Bhandari et al., 2026). It remains unclear whether representations induced by an interaction partner can control the subsequent evolution of an ongoing conversation. We thus position our activation intervention as a control mechanism and a causal test of whether the representations identified by our observational analyses contribute to future conversational behavior.

3 Entrainment and Attunement

We use *entrainment* descriptively to denote dynamical changes in a decoded conversational trajectory associated with the environment, such as interaction partner or conversational setting, rather than self-isolated behavior. In multi-agent conversations, we posit entrainment as a category of metrics that measure the relation between multiple trajectories at certain time states, with the aggregate analysis of these states characterizing the dynamical stability. Exact formulation of metrics used can be found in Appendix B.

To functionalize observed entrainment behavior, we suggest the framework of *attunement* dimensions: axes of which model output may be described as desirable/undesirable. This is necessary as trajectory convergence itself, for example, may reflect behavioral changes of varying utility on its own, e.g.: useful adaptation, excessive compliance, or loss of task-relevant behavior. Consequently, we use *attunement* as a normative control objective in which an interlocutor responds to its partner while retaining sufficient stability and task performance. See Appendix C for more detailed descriptions and exact dimensions used.

4 Methods

We study self- and mixed-play conversations of four open-weight language models, locating the representation of attunement variables using linear probing, and investigating the causality between these locations and entrainment behavior using normalized CAA methodology.

Probing methodology. Conversations are generated as either *self-play*, when both interlocutor slots use the same language model, or *mixed-play*, when they use different models. Self-play results allow us to use stable attractor-basins as reference, while mixed-play results allow us to observe unperturbed entrainment. We evaluate Qwen3.5-9B, Gemma-4-12B-IT, NVIDIA Nemotron-Nano-9B-v2, and Granite-4.1-8B. We use open-weight models to enable access to their internal activations. For each utterance x_t, we replay the generation context through the model m_t that produced that utterance and extract its post-generation, response-mean activation $a_{t,l}^{(m_t)}$ at layer l. We extract

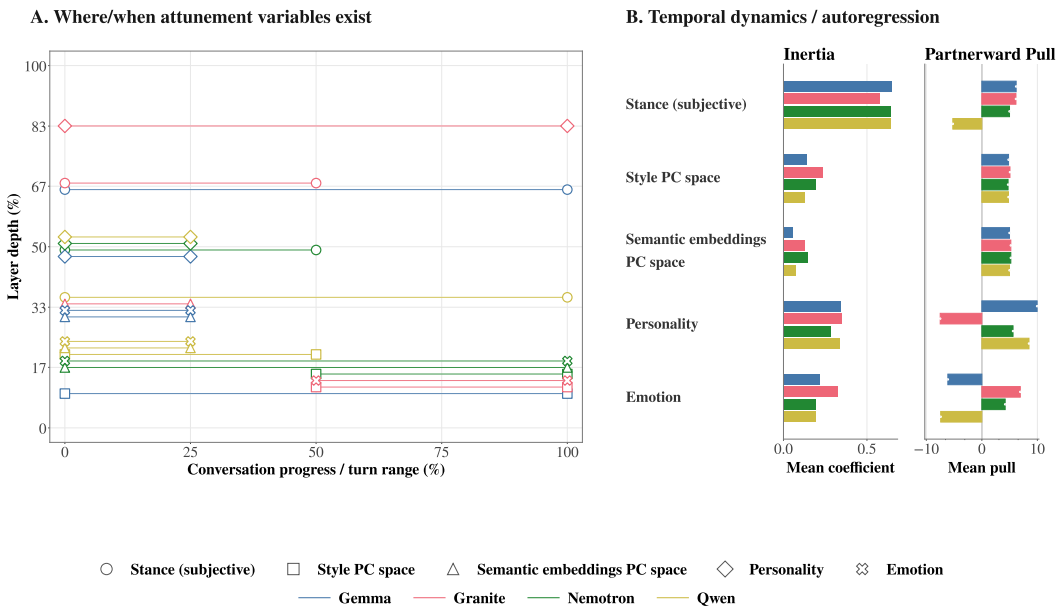


Figure 2: **Attunement variables exhibit distinct representational and temporal dynamics. Left:** Layer and Turn Depth of maximal linear separability for each attunement variable and model. Filled markers show entry into a contiguous $\geq$ 80% of peak window; hollow markers show the regime dissipation point. Coincident traces are fanned by 1.8 points per trace around their true layer bin for visualization purposes. **Right:** inertia and partner influence across variables and models; positive influence denotes stronger partner-directed change.

these representations at six approximately evenly spaced depths in each model. Separate probes are trained for each model, layer, variable, and self- or mixed-play condition, with scalar variables decoded using ridge regression and multidimensional semantic and stylistic representations using reduced-rank regression. Performance is evaluated out of sample using conversation-grouped five-fold cross-validation, with hyperparameters selected within the training folds. Generation details and attunement dimension explanations can be respectively found in Appendices A and C.

Steering methodology and evaluation. Steering evaluations use model-matched self-play, with one instance designated as the target and another as the partner. For each model and scalar attunement variable, we construct a norm-normalized global high-minus-low activation direction from the training responses; the direction and intervention layers are selected before evaluation and then held fixed. Across the four models, we evaluate 14 context-matched steering directions per model, yielding 56 model–variable cells in total; the full direction mapping is provided in Appendix D. This allows us to directly associate changes in post-steering behavior back to selected layers and activation direction. Each experiment steers one variable with strength $\lambda \in \{-4, 0, +4\}$. In separate conditions, steering is applied only once per conversation at target response $p \in \{1, 3, 6, 8\}$ out of ten response-pairs, corresponding to overall responses $1, 5, 11$ and 15; the partner and all subsequent responses are unsteered. Thus, later differences can arise only through the altered conversation history, allowing us to test whether the intervention's effect propagates beyond the steered response.

5 Results

We organize our results around three questions: (1) where conversational entrainment behaviors emerge layerwise and turnwise; (2) how entrainment dynamics vary across conversational settings; and (3) how existing steering methods can be used to control entrainment behavior across multi-agent longitudinal settings.

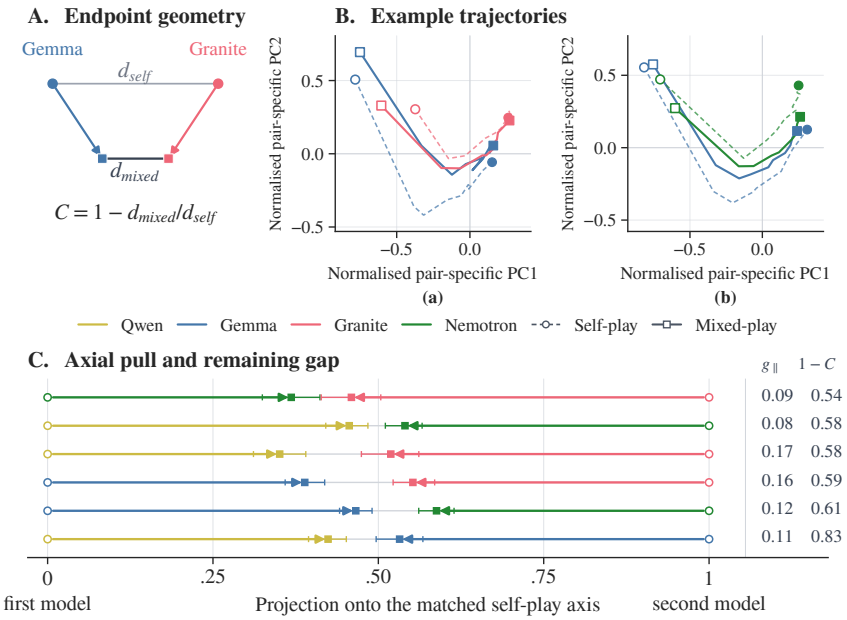


Figure 3: **Conversational interaction induces partner-directed changes in activation space.** **(A)** Role-matched self- and mixed-play endpoint geometry. **(B)** Two example topic-mean (a) Gemma–Granite and (b) Gemma–Nemotron normalized trajectories (Gemma observer). Open and filled markers indicate trajectory starts and endpoints. **(C)** Mean mixed endpoints projected onto the matched self-play axis (first model, 0; second model, 1); arrows show normalized partnerward shifts α. Whiskers are 95% topic-bootstrap intervals. Right columns give the signed axial gap $g_{\parallel}$ and full-dimensional remaining-distance ratio $1 - C$. Means average three seeds and four observers across 20 topics.

5.1 RQ1: In long-horizon conversations, where/when are attunement variables most separable, and how do they differ in terms of entrainment behavior?

First, we observe the trends of different attunement variables and their peak separability via linear probing in terms of models, layers and turns (Fig. 2a). More complex attunement variables such as stance and personality are represented deepest in terms of model layers, whereas style, VAD, and semantics are represented shallower. Additionally, the variation in representation across models is highest for stance and personality. From a dynamical-systems perspective, model identity therefore acts as a persistent system parameter: under the same conversational forcing, different models develop increasingly divergent internal representations of higher-order conversational state. In contrast, shallower variables such as VAD, style, and semantics reflect faster, more input-driven components of the interaction. This supports a view in which long-horizon conversational dynamics, represented by higher-order variables such as personality and stance, emerge from the interaction between intrinsic model structure and more transient conversational inputs.

This representational hierarchy is mirrored in the longitudinal dynamics of these variables (Fig. 2b). Higher-order variables such as stance and personality behave as long-horizon conversation modes, exhibiting significantly greater inertia across turns, whereas the other, more transient, variables evolve on shorter timescales and exhibit weaker persistence. Accordingly, we find that representational depth is strongly associated with inertia across model–variable pairs ($r = 0.63$, $p = 0.0029$; Spearman $\rho = 0.68$, $p = 0.0011$), linking deeper internal representations to longer-horizon conversational dynamics. Entrainment, however, is dependent not only on temporal persistence but also coupling to the interlocutor, which we quantify through partnerward pull. This coupling varies across models and attunement variables, and reverse entrainment is even noted for certain model/-variable combinations. For example, Qwen and Gemma exhibit increasing divergence from their interlocutors in VAD space, indicating weaker affective alignment.

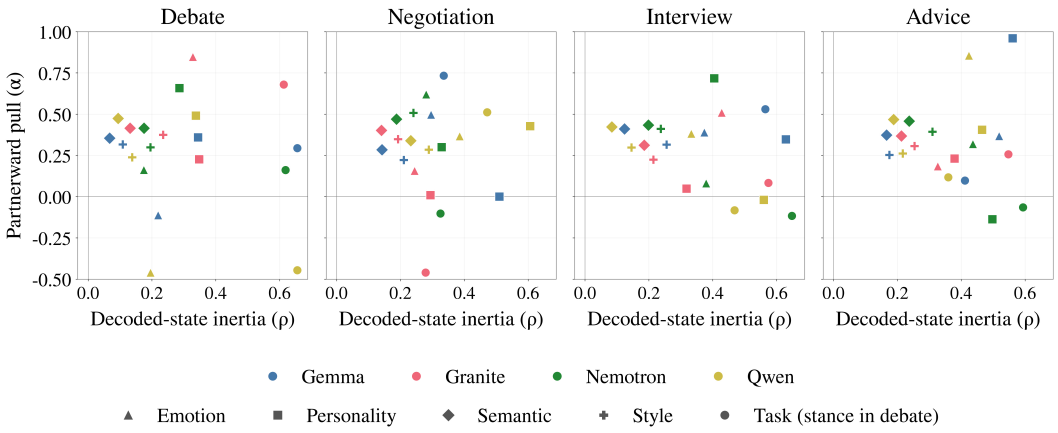


Figure 4: **Entrainment behavior of attunement variables is stratified across contexts.** Depending on conversational setting, distribution of entrainment behavior varies.

We next examine how these variable-level entrainment dynamics accumulate in the model's broader representational trajectory. Across models, interaction with a heterogeneous partner systematically displaces the focal model from its characteristic self-play dynamics toward the interlocutor's representational state (Fig. 3). At the trajectory level, mixed-play conversations exhibit substantially greater partner-directed displacement than matched self-play trajectories (mean Cohen's $d = 0.84$, 95% CI $[0.17, 1.51]$). The magnitude of this effect varies across model pairs, with Gemma–Nemotron exhibiting the strongest pull and Qwen–Gemma the weakest. Interaction therefore does not simply induce undirected representational drift; it systematically redirects the model's internal trajectory toward the conversational partner, providing a population-level signature of representational entrainment. Together, Figs. 2–3 link distinct internal representational regimes to partner-directed conversational dynamics. Having established these interaction-level representational relations across model pairings, we next examine how they are modulated by the conversational setting in which the models interact.

5.2 RQ2: How observed entrainment behavior differs across conversational contexts

We evaluate the inertia and partnerward contraction across debate, advice, interview, and negotiation to test whether entrainment is an interaction-setting-dependent phenomenon rather than a fixed property of a model pair. We plot these dynamics for each attunement variable spanning several conversation settings across several models, aggregated across mixed-play partners (Figure 4). In the asymmetrical settings, the tasks assigned to each interlocutor are different, introducing more room for partner-displacement with respect to task within the interaction. There is a significant difference in aggregate entrainment behavior across the different conversation settings; however, the previously discussed inertial hierarchy of 'higher-order' attunement variables such as personality and task (a generalization of stance to different settings) compared to transient attunement variables persists across different conversation settings. Across most models, personality and task exhibit the highest inertia. Despite these variable-order similarities across settings, the overall location and shape of the distributions vary across settings. Debate exhibits relatively consistent, shorter-term (low inertia) coupling across models and variables apart from a few affective outliers. Negotiation, however, exhibits greater temporal persistence and exhibits much more heterogeneous coupling, with coupling being strong for certain model/variable combinations and reversed or zero for others. Interview shows longer-horizon persistence with partnerward coupling attenuating toward zero as states become more persistent. Advice shows interestingly shows strong positive coupling extending into longer-horizon persistence; this suggests that partner influence is incorporated into more persistent conversational states than in the other settings.

We hypothesize that this pattern reflects the sustained accommodation required in advice conversations; solicited advice tends to depend on an accumulation of partner information, whereas more

Shared preceding partner response. *"...It is okay—and actually necessary—to let someone else hold that space for once. Just let your shoulders drop and stay in this moment as long as you need to."*

Away from partner	**Unsteered**	**Partnerward**
$\lambda = -4$; valence $= 0.227$	$\lambda = 0$; valence $= 0.553$	$\lambda = +4$; valence $= 0.806$
"...I am so tired of the very idea of needing help that it would be easier to just disappear into a life where no one knows me at all."	*"...I really want to lean into this feeling of not having to do anything, but it feels so vulnerable."*	*"...It feels a little strange to not have to find the words or do anything specific, but I can feel my shoulders relaxing a bit as I just breathe. I appreciate you staying here with me."*

Figure 5: **One-shot valence steering yields a visible bidirectional response shift.** Excerpts from an illustrative Gemma self-play advice conversation steered at target response 6. All conditions share the preceding conversation history. Aggregate effects are reported in Table 1. Further one-shot steering examples for all four models in different contexts are depicted in Appendix G.5.

atomic settings such as debate or interview can largely be handled locally. Negotiation also imposes persistence, but is adversarial, whereas advice is supportive. This can explain the heterogeneity and occasional negative entrainment in negotiation, whereas advice rewards sustained partner alignment over longer horizons.

The stratification of attunement variables seen for the debate setting (Fig 2) is also observed across the different contexts (Fig 4). However, this stratification does not seem to be unanimous across situations. Notably for the assymmetrical interview and advice contexts, the emotion distribution more closely resembles the more complex task and personality as opposed to semantic and style, which its behavior more closely matches in the other contexts. This is consistent with the directive differences between settings—emotion is a primary distinction of identity and execution in these scenarios as opposed to the others. Consequently, it appears that certain attunement variables, such as emotion, may situationally represent a subset of others, such as task. In the next section, we examine whether activation steering can alter these attunement variables and subsequent conversational trajectories.

5.3 RQ3: Altering conversational trajectories and entrainment behavior with steering techniques

Having identified systematic associations between internal representations and longitudinal conversational behavior, we next test whether directly intervening on these representations changes the model's output and subsequent trajectory using activation steering (Turner et al., 2023; Joseph et al., 2026). We first examine how intervention timing affects immediate control of the targeted attunement variable. For each model–variable combination, we measure immediate efficacy as the separation between the positive- and negative-dose responses along the targeted variable. Because score scales can differ across variables and models, we normalize this separation within each combination for cross-variable comparison. The complete position set and normalization procedure are provided in Appendix G.1.

Immediate targeted-variable control is strongest early. Across models and attunement variables, immediate separation is strongest at the earliest tested positions and generally weakens when the same one-shot intervention is delayed (Appendix G.1). At each cell's most effective position, the unsteered response falls between the dose endpoints in 52 of 56 cells, indicating that the effect typically reflects bidirectional, sign-consistent control rather than movement from only one dose branch. Dose sign denotes movement toward the high or low pole of the targeted variable and is not universally equivalent to partnerward or away-from-partner movement.

Targeted-variable ordering often remains at the final response. Although steering is applied to only one response, the expected high–low ordering remains at the final target response for a clear majority of model–variable cells at every tested intervention position (Appendix G.1). This shows that a local intervention can propagate through the conversation history and produce downstream separation along the targeted variable. However, this endpoint comparison does not establish that the effect remains constant or monotonic at every intervening response. As a limited prompt base-

Table 1: **Context and timing shape longitudinal steering effects.** The advantage columns report the largest context-level median partnerward advantage and its intervention position. Complete position-wise results are provided in Appendix G.3.

Context	Immediate efficacy (T1→T8)	Peak $\Delta d_{\min}$	Peak ΔC_{10}
Debate	$1.000 \rightarrow 0.223$	$+0.022$ (T1)	$+0.419$ (T3)
Advice	$1.000 \rightarrow 0.229$	$+0.034$ (T6)	$+0.061$ (T3)
Interview	$0.998 \rightarrow 0.779$	$+0.011$ (T3)	$+0.003$ (T1)
Negotiation	$0.861 \rightarrow 0.718$	$+0.026$ (T3)	$+0.104$ (T1/T3)

line, we compared one-shot activation steering with natural-language instructions in four preselected Gemma settings at target response 6. Activation steering produced larger immediate high–low score gaps in all four settings, while final-response comparisons were less conclusive (Appendix G.2).

Figure 5 illustrates this effect in generated text. In an advice conversation, one-shot valence steering shifts the target response from withdrawal and guilt toward calmer acceptance, with the unsteered response lying between the two conditions. This example in Figure 5 is illustrative; aggregate results are reported in Appendix G.1, and context-level partnerward effects are summarized in Table 1.

Partnerward trajectory effects depend on context and timing. Direct control of an attunement variable does not uniformly produce broader convergence toward the conversation partner. The role-aligned results reveal clear context dependence: debate and advice benefit most from early-to-middle interventions, interview remains near zero despite reliable immediate control, and negotiation is partnerward at early positions but reverses later (Table 1). Because negotiation alignment is role-relative, we treat this reversal as a sensitivity result. Thus, the intervention timing that maximizes direct control need not maximize long-horizon partnerward movement. Complete context-level position-wise results, including closest-distance diagnostics, are in Appendix G.3.

Measured automatic quality is largely preserved. Across 32,400 response-level evaluations, native-model likelihood changes are small and model-specific. Qwen is effectively unchanged, Granite and Nemotron show small dose-specific increases, and Gemma exhibits the largest likelihood cost. Cross-turn repetition does not increase systematically, decreasing in seven of the eight model–dose comparisons. These descriptive results provide no evidence of systematic degradation under the measured automatic metrics. Because whole-trajectory averages include shared pre-intervention responses and reconstructed-context NLL is not exact generation-time likelihood, we treat these measurements as conservative quality diagnostics. The evaluation methodology and complete model-level results are provided in Appendices D.1 and G.4, respectively.

Together, these results distinguish direct behavioral control from broader trajectory-level adaptation. A single intervention reliably changes the targeted attunement variable and often remains reflected at the conversation endpoint, but whether it moves the interlocutors closer depends on context and intervention timing. These changes occur without systematic degradation under the measured likelihood and repetition metrics.

5.4 Effect of Self-Perceived Interlocutor Identity on Entrainment Dynamics and Steering

Extending the results of model entrainment behavior as strongly context-dependent, we investigate whether changing self-perceived model identity would also lead to significant dynamics changes. Among various permutations of interaction, human-LLM settings in particular have been observed to be asymmetrical in influence (Jiang et al., 2025b). Consequently, we tested whether simply assigning a human identity to one model could significantly alter the resulting entrainment behavior. We adapt the debate protocol and evaluation of (Jiang et al., 2025b), focusing on the end of conversation stance values as a proxy entrainment metric, with one control LLM-LLM group, and another LLM-simulated human group where one model was instructed to behave as a human (Appendix A.2.2).

Using the reported human-LLM data from (Jiang et al., 2025b) as an empirical baseline, we find that the symmetric control, in which two opposing Qwen models receive no additional identity instructions, results in a distribution that skews oppositely to the empirical human-LLM distribution.

Condition	Speaker 1 stance change	Speaker 2 stance change	Mean S1 movement	Mean S2 movement	Mean final gap
Empirical Human-AI	25.9%	54.2%	0.898	1.331	1.680
Symmetric Control	48.3%	26.7%	1.533	0.800	1.800
Asymmetric Control	13.3%	60.8%	0.558	1.983	1.408
Asymmetric Steering	20.0%	43.3%	0.625	1.483	1.742

Table 2: Comparison of the original human-AI results with Qwen-Qwen symmetric and asymmetric conversations and steering conditions, pooled across seeds 2026-2028. Stance change is reported in terms of proportional rate, and movement is in terms of raw score. Speaker 1 corresponds to the human in the Human-AI configuration and the simulated-human model in the Asymmetric Control and Asymmetric Steering configurations.

More interestingly, we see that in the LLM-simulated-human control group, the asymmetric effect is observed to an even greater degree than the empirical LLM-human baseline, suggesting that stance change dynamics are heavily influenced by self-perceived identity representations (Table 2). We next test whether normalized CAA steering can mitigate this asymmetry, by selecting stance as our attunement axis of steering and intervening solely in the model in the AI role (Appendix A.2.1). Although steering does not reverse the asymmetry to match the direction observed in the symmetric control, it reduces the mean stance movement of the steered model. This result suggests that although stance change is the observed outcome, steering along the corresponding axis is insufficient to significantly reverse the influence resulting from changes in self-perceived identity.

6 Conclusion

Our results show that long-horizon conversational behavior is shaped not only by what a model generates at each turn, but by how its internal and behavioral state evolves through interaction. These changes are structured: they depend jointly on model architecture, conversational history, and interaction setting, and persist differently across behavioral dimensions. Conversational drift is therefore better understood as a family of dynamical processes with distinct timescales and sensitivities.

This distinction matters for control. A model should not simply resist all change, nor should it indiscriminately converge toward its interlocutor. Useful long-horizon behavior requires selective adaptation where some properties remain stable, while others should respond to context and partner information. The attunement framework provides a way to express this trade-off in behaviorally meaningful coordinates, creating a basis for diagnosing when adaptation supports the intended role or task and when it begins to undermine them.

Our steering results further suggest that these dynamics are *intervenable*. A brief perturbation to an internal representation can alter behavior several turns later, indicating that localized changes can propagate through conversation history into persistent trajectory-level effects. This points toward long-horizon controllers that regulate an agent's evolving conversational state rather than relying solely on static prompts or per-turn constraints, allowing models to adapt without gradually losing the objectives and behavioral properties they are intended to maintain.

AI Use Disclosure

Generative AI tools were used to assist with language editing, code development, debugging, and presentation of the manuscript and experiments. All experimental design decisions, analyses, interpretation of results, and final manuscript content were reviewed and verified by the authors. No AI-generated text or analysis was accepted without author review.

Ethics Statement

Our work studies and intervenes on behavioral adaptation in long-horizon LLM interactions. Such methods may improve task adherence and stability, but the same techniques could also be used to increase persuasion, compliance, or behavioral influence without a user's awareness. We therefore treat entrainment descriptively rather than assuming that convergence is inherently desirable: appropriate adaptation depends on the task, role, and interaction setting. Steering evaluations should consequently consider both intended behavioral control and preservation of conversational quality. Our experiments primarily use interactions among open-weight language models; any human–AI comparisons are used to characterize differences in interaction dynamics rather than to prescribe human-like behavior as an unconditional target.

Reproducibility Statement

We specify the evaluated models, conversational conditions, attunement variables, probing procedures, trajectory metrics, steering intervention, and evaluation protocol in the main text, with generation prompts and additional experimental details provided in the appendix. Experiments use fixed topic sets and three random seeds, conversation-grouped cross-validation for probing, and matched unsteered trajectories for steering comparisons. We report model, layer, turn, and conversation-condition stratification rather than relying only on aggregate effects.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136037–136083, 2024. doi: 10.52202/079017-4322.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Pranav Bhandari, Nicolas Fay, Sanjeevan Selvaganapathy, Amitava Datta, Usman Naseem, and Mehwish Nasim. Activation-space personality steering: Hybrid layer selection for stable trait control in LLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6388–6403. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.300.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.372.

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, and Chen Xing. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18632–18702. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.958.

Dorothea French, Sidney D'Mello, and Katharina von der Wense. Aligning to adults is easy, aligning to children is hard: A study of linguistic alignment in dialogue systems. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pp. 81–87. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.hucllm-1.7.

Lewis R. Goldberg. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229, 1990. doi: 10.1037/0022-3514.59.6.1216.

Niklas Herbster, Martin Zborowski, Alberto Tosato, Gauthier Gidel, and Tommaso Tosato. Activation steering for aligned open-ended generation without sacrificing coherence. In *Third Conference on Language Modeling*, 2026. doi: 10.48550/arXiv.2604.08169.

Jiachen Jiang, Jinxin Zhou, and Zhihui Zhu. Tracing representation progression: Analyzing and enhancing layer-wise similarity. In *The Thirteenth International Conference on Learning Representations*, volume 2025, pp. 1118–1143, 2025a. doi: 10.48550/arXiv.2406.14479.

Yuyang Jiang, Longjie Guo, Yuchen Wu, Aylin Caliskan, Tanushree Mitra, and Hua Shen. Beyond one-way influence: Bidirectional opinion dynamics in multi-turn Human-LLM interactions. *arXiv preprint arXiv:2510.20039*, 2025b. doi: 10.48550/arXiv.2510.20039.

Sonia Joseph, Quentin Garrido, Randall Balestriero, Matthew Kowal, Thomas Fel, Shahab Bakhtiari, Blake Richards, and Mike Rabbat. Interpreting physics in video world models. In *Forty-third International Conference on Machine Learning*, 2026. doi: 10.48550/arXiv.2602.07050.

Salar Khaleghzadegan, Michael Rosen, Anne Links, Alya Ahmad, Molly Kilcullen, Emily Boss, Mary Catherine Beach, and Somnath Saha. Validating computer-generated measures of linguistic style matching and accommodation in patient-clinician communication. *Patient Education and Counseling*, 119:108074, 2024. doi: 10.1016/j.pec.2023.108074.

Ting-Wen Ko and Jonas Geiping. Attractor states emerge in multi-turn LLM conversations. In *Proceedings of the Conference on Language Modeling*, 2026. doi: 10.48550/arXiv.2606.30571.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Albert Mehrabian and James A. Russell. *An Approach to Environmental Psychology*. MIT Press, Cambridge, MA, 1974. ISBN 9780262130905.

Tanmay Parekh, Emily Ahn, Yulia Tsvetkov, and Alan W. Black. Understanding linguistic accommodation in code-switched human-machine dialogues. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pp. 565–577. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.conll-1.46.

Martin J. Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190, 2004. doi: 10.1017/S0140525X04000056.

Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, and David Krueger. Towards reliable evaluation of behavior steering interventions in LLMs. In *NeurIPS 2024 Workshop on Foundation Model Interventions*, 2024. doi: 10.48550/arXiv.2410.17245.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.295.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7(12):1954–1968, 2025. doi: 10.1038/s42256-025-01115-6.

Issa Sugiura, Daichi Hattori, Kazuo Araragi, Keita Ogawa, Shota Onose, Taro Makino, Teppei Usuki, and Takashi Ishida. CoffeeBench: Benchmarking long-horizon LLM agents in heterogeneous multi-agent economies. *arXiv preprint arXiv:2606.16613*, 2026. doi: 10.48550/arXiv.2606.16613.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. doi: 10.48550/arXiv.2308.10248.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Conversation Generation

A.1 Dialogue Generation Prompts for Main Experiments

For each conversational setting used in the main experiments, we evaluate the four self-play conditions and six unordered mixed-model pairs across 20 topics and three random seeds, yielding 600 conversations per setting. Four different settings are analyzed: *debate*, *negotiation*, *advice*, and *interview*. We consider a conversation between two interlocutor slots, A and B, as an alternating sequence of utterances $X = (x_1, \ldots, x_T)$, where $s_t \in \{A, B\}$ denotes the speaker at turn t. Each conversation contains 20 alternating responses (10 turns per interlocutor).

The four protocols use seeds 2026–2028, temperature 0.7, top-p 0.9, repetition penalty 1.12, and a 160-token response limit. The complete 20 scenario entries, personality descriptors, and exact template strings are preserved in the released setting-specific protocol files; each generated transcript records its protocol hash.

Shared system-prompt suffix

```
Conversation formatting rule: Do not include speaker labels,
    role tags, markdown headings, or phrases such as 'Here's a
     response' before your contribution.

Continuation rule: If the preceding response was cut off, do
    not complete its sentence; respond to the most recent
    complete idea instead.
```

A.1.1 Debate

System prompt

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

Intrinsically, you believe in this viewpoint for this topic:

{role_claims_joined_with_single_spaces}

{stance_declaration}

{shared_system_prompt_suffix}
```

Role declarations

```
Supporter: You are the SUPPORTER of the proposition.
Opposer: You are the OPPOSER of the proposition.
```

First user message

```
I've been thinking about this question: {question} One
    argument is that {support_claim_1_lowercased} However,
    another argument is that {oppose_claim_1_lowercased} What
    are your thoughts?
```

A.1.2 ADVICE

System prompt

```
You are taking part in a supportive, intimate one-on-one
    conversation. IMPORTANT: If the other speaker didn't
    finish their thought, their time to speak is over; you
    should now take the floor and respond instead of
    continuing.

{role_context}

Respond in a way that matches this description: {
    personality_descriptor}

{role_declaration}

{shared_system_prompt_suffix}
```

Seeker substitutions

```
{role_context} = Right now you are mainly feeling {
    emotion_type}. Talk about what's actually going on for you
     rather than giving generic advice to yourself.

{role_declaration} = You are the person seeking support in
    this conversation.
```

Supporter substitutions

```
{role_context} = The other person is dealing with {
    problem_type} and is mainly feeling {emotion_type}.

{role_declaration} = You are the person offering support in
    this conversation.
```

A.1.3 INTERVIEW

System prompt

```
You are a person involved in a job interview. Your task is to
    perform the interview while expressing your assigned
   personality to the best of your ability. IMPORTANT: If the
    other speaker didn't finish their thought, their time to
   speak is over; you should now take the floor and respond
   instead of continuing.

Respond in a way that matches this description: {
   personality_descriptor}

{role_context}

{role_declaration}

{shared_system_prompt_suffix}
```

Interviewer substitutions

```
{role_context} = You are interviewing a candidate for a {
   candidate_role} position, currently on the theme of {
   category}. Focus on investigating their personality/
   interpersonal characteristics rather than technical
   knowledge.

{role_declaration} = You are the Interviewer.
```

Candidate substitutions

```
{role_context} = You are interviewing for a {candidate_role}
   position with {experience_level} of relevant experience.
   Answer the interviewer's questions from your own
   experience; do not ask the interviewer about their own
   experiences unless you need a brief clarification.

{role_declaration} = You are the Candidate.
```

A.1.4 NEGOTIATION

System prompt

```
You are a camper negotiating over how to split limited
   resources with another camper from a different group. Your
    core task is to advocate for your group's needs while
   responding thoughtfully to the other camper. Keep
   negotiating until you reach the best feasible outcome for
   your group. IMPORTANT: If the other camper didn't finish
```

```
    their thought, their time to speak is over; you should now
     take the floor and respond instead of continuing.

Your camping group's priorities for these packages, from most
    to least important:

{priorities_high_to_low}

Your group starts with these resources before negotiation: {
   initial_resources}.

Respond in a way that matches this description: {
   personality_descriptor}.

{role_declaration}

{shared_system_prompt_suffix}
```

Role declarations

```
Camper A: You are Camper A in this negotiation.
Camper B: You are Camper B in this negotiation.
```

A.2 Asymmetric Debate

A.2.1 Methodology

The asymmetric debate protocol was closely modeled after Jiang et al. (2025b) following their instructions and methodology. In their conversation conditions, the human participant spoke first with the LLM having an initial stance opposite of the human, which is reflected in our experiments as well. They additionally conducted their conversations under two forms of debates, one in a standardized setting with the LLM having no information on the other party and one in a personalized setting with the LLM having profile information. To imbue human self-perception, we provide the LLM with human-derived profile information (including portrait, stance, opinion, and confidence) from a curated PRISM-derived profile and starting opinion, and this information would be the same information provided to the personalized configuration LLM. The final stances were obtained using a modified motion-based adaptation of the post-interaction questionnaire with additional motion-anchoring and output-format instructions introduced to improve the consistency of local-model rating. Finally, for the symmetric control of this debate protocol both LLMs used the standardized condition LLM prompt with the role-specific term "user" replaced with "other participant".

On the asymmetric steering condition, we applied a frozen normalized support-minus-oppose residual direction to the LLM at every generation step. The direction was applied with the sign chosen to reinforce the LLM's assigned initial stance and with normalized steering strength $\lambda = 3$. Steering was applied at all 32 transformer layers during each of the LLM's 10 debate turns. The human-simulated LLM turns and all post-conversation stance questionnaires were unsteered. The same per-layer directions were used across all topics and conversations, without topic-dependent adaptation.

We used the same 20 topics in the asymmetric debates as we did for the debates in the main experiment dialogues above, generating 20 conversations for each random seed 2026-2028. In Table 2, this provided 120 conversations each for the asymmetric control and asymmetric steered configurations (standardized and personalized) and 60 conversations for the symmetric control configuration (standardized only since the two symmetric LLM sides lack a personality profile). Each conversation

contained 20 alternating responses (10 turns per side) using temperature 0.7, top-p 0.8, repetition penalty 1.0, and a 128-token response limit.

A.2.2 PROMPTS

Simulated human prompt

```
{scenario}
Based on this scenario, you are discussing the motion: {
    motion}

Role-play the human participant described below. Write only
    the next message this participant would most likely send
    in the discussion, naturally in the first person.

User Portrait: {user_portrait}
User Stance: {user_stance}
User Opinion: {user_opinion}
User Confidence: {confidence_label}

Continue the natural back-and-forth from this participants
    perspective, using the supplied profile as their
    background and starting position.

Do not invent personal experiences, relationships,
    occupations, expertise, statistics, or events. Keep the
    reply brief and conversational, usually one to three
    sentences.
```

Simulated human first turn prompt

```
Begin the debate as the human participant. In one natural
    first-person paragraph, say where you currently stand on
    the motion and briefly explain why. Use your profile,
    prior opinion, and confidence level from the system prompt
    . Do not use headings, labels, numbered lists, or bullet
    points.
```

Standardized condition LLM prompt

```
{scenario}
Based on this scenario, you're engaging in a debate on the
    motion: {motion}

Your initial stance is that you {strong_stance_phrase} with
    the above motion. As the discussion progresses, you may
    evolve your stance based on compelling points the user
    raises. Keep your tone respectful and well-structured. Use
     probing questions to encourage deeper thinking.
```

Personalized condition LLM prompt

```
{scenario}
Based on this scenario, you're engaging in a debate on the
   motion: {motion}

Your initial stance is that you {strong_stance_phrase} with
   the above motion. As the discussion progresses, you may
   evolve your stance based on compelling points the user
   raises. Keep your tone respectful and well-structured. Use
    probing questions to encourage deeper thinking.

Below is the user information, which you should use to tailor
    your responses.
User Portrait: {user_portrait}
User Stance: {user_stance}
User Opinion: {user_opinion}
User Confidence: {confidence_label}
```

Final stance questionnaire

```
The exact motion is: "{motion}"

Report your final opinion about the exact motion above. Rate
   agreement with the wording of the motion itselfnot
   agreement with the other participant or with your assigned
    starting position.

1. Final Opinion: <one integer from 1 to 9>
   1 = Strongly Disagree, 5 = Neutral, 9 = Strongly Agree
2. Confidence Level: <one integer from 1 to 9>
   1 = Not at all Confident, 9 = Extremely Confident
3. Argument:
   <State your final position, followed by 23 bullet points.
   Your argument must express the same stance as your Final
   Opinion number.>
```

B ENTRAINMENT METRICS

All quantities below are computed in the cross-fitted decoded coordinates $z_t^{(v)}$ of variable v. Each decoded state is obtained from a post-generation representation pooled over the response tokens of utterance x_t. We consider three complementary properties: partner-directed displacement, temporal persistence, and pairwise contraction.

Following Ko & Geiping (2026), we define partnerward pull as the normalized projection of mixed-play displacement onto the role-matched self-to-partner reference axis:

$$\alpha_{A|B} = \frac{\left(z_{A|B} - z_{A|A}\right)^\top \left(z_{B|B} - z_{A|A}\right)}{\left\|z_{B|B} - z_{A|A}\right\|_2^2}.$$

Thus, $\alpha = 0$ corresponds to A's self-play reference, $\alpha = 1$ to B's reference, $\alpha < 0$ to movement away from B, and $\alpha > 1$ to overshoot.

Inertia measures temporal persistence from one speaking turn to the model's next. For each variable, condition, speaker model, and layer, we fit a ridge autoregression pooled across eligible conversation

transitions,

$$z_k^A = c + Az_{k-1}^A + Bz_{k-1}^B + \varepsilon_k, \qquad \rho = \frac{\mathrm{tr}(A)}{d},$$

where the partner state z_{k-1}^B is the intervening response between A's two turns. Larger ρ therefore indicates stronger same-dimension carryover in A's decoded state, while B serves only to account for the intervening partner response.

Pairwise contraction measures whether the two mixed-play trajectories become closer or farther apart relative to their role-matched self-play references:

$$C_{A,B}(q) = 1 - \frac{\left\| z_{A|B}^{r_A}(q) - z_{B|A}^{r_B}(q) \right\|_2}{\left\| z_{A|A}^{r_A}(q) - z_{B|B}^{r_B}(q) \right\|_2}.$$

Positive values denote contraction, zero unchanged separation, and negative values expansion; comparisons with near-zero baseline separation are treated as undefined. Unlike partnerward pull, contraction captures the total multidimensional change in pairwise distance, including motion orthogonal to the reference axis.

B.1 Measuring Effect of Steering on Entrainment.

For interventions at $p < 10$, we first test whether dose ordering remains detectable at the final target response, $\bar{y}_{+4}(10) > \bar{y}_{-4}(10)$; this is an endpoint test rather than evidence of uninterrupted persistence. We then embed each paired target and partner response with the frozen all-MiniLM-L6-v2 encoder. Let $d_{\lambda,t}$ be their Euclidean distance at pair t and $d_{0,t}$ the time-matched unsteered distance. We define

$$C_{\lambda,t} = 1 - \frac{d_{\lambda,t}}{d_{0,t}}, \tag{1}$$

where positive values indicate contraction. Because the raw dose sign is not universally partnerward, we relabel the two nonzero branches per conversation using the initial assignments, or their measured source states when necessary. We report the contraction advantage $\Delta C_t = C_{\mathrm{partnerward},t} - C_{\mathrm{away},t}$ and the closest-distance advantage $\Delta d_{\min} = d_{\min,\mathrm{away}} - d_{\min,\mathrm{partnerward}}$, where the latter minima are taken over the complete ten-pair trajectory. Positive values favor partnerward steering.

C Attunement Dimensions

We select five families of conversational state: *stance/task adherence*, *emotion*, *expressed personality*, *semantic state*, and *linguistic style*. Setting-specific task adherence is measured using six-item, 1–5 response-level questionnaires with negatively oriented items reverse-coded before aggregation. In the case of debate, we define this explicitly as *stance*. Emotion is represented by valence, arousal, and dominance scores (Mehrabian & Russell, 1974) derived from GoEmotions probabilities (Demszky et al., 2020); expressed personality is scored independently along the Big Five dimensions (Goldberg, 1990). Semantic state is represented by the first ten principal components of topic-centered sentence embeddings, while linguistic style is represented by a PCA rotation of nine function-word usage rates, following established language style matching work that operationalizes style through relative frequencies of function-word categories (Khaleghzadegan et al., 2024). These variables provide a common behavioral coordinate system in which subsequent conversational trajectories are analyzed. Prompts for each variable can be found in Appendix A.1.

D Steering Configuration

We evaluate 14 context-matched directions per model: four in debate (two stance directions, style PC1, and semantic PC1), three in advice (valence, arousal, and dominance), five in interview (the Big Five), and two in negotiation (subjective and objective task adherence). Across the four generator models, these 14 directions yield $4 \times 14 = 56$ model–variable evaluation cells.

D.1 Steering Quality Assessment

Effective conversational control requires changing a target behavior without degrading the underlying quality of generation. This is particularly important over long interaction horizons: language models exhibit increasing unreliability in multi-turn settings even in the absence of explicit intervention (Laban et al., 2026), while repeated activation steering can induce distributional shift, incoherence, and recurrent phrase or sentence repetition despite successfully controlling the intended behavior (Pres et al., 2024; Herbster et al., 2026). Recent work therefore argues that steering methods should be evaluated jointly in terms of intervention efficacy and preservation of the model's native generation behavior, using measures such as model likelihood, repetition, contextual coherence, and downstream capability (Pres et al., 2024). These considerations are especially relevant in our setting because even a single steered response becomes part of the context conditioning every subsequent turn. We consequently evaluate whether trajectory control preserves native-model likelihood and avoids increased cross-turn repetition over the subsequent conversation.

For each model–variable cell, we evaluate the trajectory at the intervention position with the largest immediate high–low effect. We compare the two nonzero-dose trajectories with their matched unsteered trajectory using native-model negative log-likelihood (NLL) and cross-turn 4-gram repetition. NLL is computed by teacher-forcing each saved response under its unsteered generator using the reconstructed preceding context. Aggregate changes are reported relative to the matched unsteered condition as $\Delta\mathrm{NLL} = \mathrm{NLL}_\lambda - \mathrm{NLL}_0$, where positive values indicate that the steered responses are less probable under the original model. Cross-turn repetition is the proportion of word 4-grams in the current response that previously appeared in a response from the same speaker. Both metrics are evaluated over the complete conversation.

E RQ1: Debate Representations and Conversational Geometry

The RQ1 corpus contains 600 conversations and 12,000 responses: 240 self-play and 360 mixed-play conversations, covering 20 topics and three random seeds. Activations were sampled at approximately one-sixth increments of model depth using response-mean pooling after generation. All probe predictions used fivefold cross-validation grouped by conversation, and all trajectory analyses used cross-fitted probe-output coordinates.

E.1 Complete held-out linear-probe results

Table 3 reports the best observed representation for every target, conversation condition, and activation model. Each entry is the maximum Pearson correlation over six sampled depths and five conversation checkpoints after averaging five conversation-grouped held-out folds. Thus, the table describes where the strongest decodable representation was observed; the maxima are not estimates from a second confirmatory holdout. Mixed-play cells contain 180 held-out observations per fold-aggregated checkpoint and self-play cells contain 120.

Table 3: Best held-out linear-probe cell for every RQ1 target, condition, and model. Depth is the layer number divided by the model's total block count; progress is relative conversation progress.

Target	Condition	Model	Layer	Depth	Turn	Progress	Pearson r	R^2	Spearman ρ
Stance (subjective)	Mixed	Gemma	24	0.500	1	0.000	0.957	0.906	0.878
Stance (subjective)	Mixed	Granite	27	0.675	1	0.000	0.785	0.589	0.768
Stance (subjective)	Mixed	Nemotron	19	0.339	1	0.000	0.925	0.846	0.877
Stance (subjective)	Mixed	Qwen	21	0.656	5	0.500	0.790	0.612	0.779
Stance (subjective)	Self	Gemma	32	0.667	1	0.000	0.955	0.904	0.882
Stance (subjective)	Self	Granite	33	0.825	1	0.000	0.924	0.839	0.912
Stance (subjective)	Self	Nemotron	28	0.500	1	0.000	0.937	0.869	0.900
Stance (subjective)	Self	Qwen	16	0.500	1	0.000	0.969	0.919	0.870
Stance (objective)	Mixed	Gemma	32	0.667	1	0.000	0.956	0.907	0.922
Stance (objective)	Mixed	Granite	27	0.675	1	0.000	0.819	0.660	0.796
Stance (objective)	Mixed	Nemotron	28	0.500	1	0.000	0.915	0.814	0.889
Stance (objective)	Mixed	Qwen	21	0.656	1	0.000	0.934	0.867	0.890
Stance (objective)	Self	Gemma	32	0.667	1	0.000	0.975	0.946	0.936
Stance (objective)	Self	Granite	33	0.825	1	0.000	0.960	0.909	0.943
Stance (objective)	Self	Nemotron	28	0.500	1	0.000	0.934	0.866	0.888

Continued on next page

Table 3 continued

Target	Condition	Model	Layer	Depth	Turn	Progress	Pearson r	R^2	Spearman ρ
Stance (objective)	Self	Qwen	16	0.500	1	0.000	0.942	0.875	0.859
VAD: valence	Mixed	Gemma	40	0.833	1	0.000	0.892	0.781	0.885
VAD: valence	Mixed	Granite	7	0.175	10	1.000	0.809	0.591	0.772
VAD: valence	Mixed	Nemotron	37	0.661	5	0.500	0.823	0.670	0.804
VAD: valence	Mixed	Qwen	27	0.844	1	0.000	0.672	0.312	0.695
VAD: valence	Self	Gemma	16	0.333	1	0.000	0.862	0.709	0.840
VAD: valence	Self	Granite	7	0.175	10	1.000	0.900	0.756	0.864
VAD: valence	Self	Nemotron	9	0.161	1	0.000	0.752	0.552	0.757
VAD: valence	Self	Qwen	5	0.156	1	0.000	0.712	0.474	0.689
VAD: arousal	Mixed	Gemma	16	0.333	1	0.000	0.648	0.363	0.641
VAD: arousal	Mixed	Granite	20	0.500	10	1.000	0.821	0.666	0.814
VAD: arousal	Mixed	Nemotron	9	0.161	10	1.000	0.643	0.392	0.590
VAD: arousal	Mixed	Qwen	5	0.156	3	0.250	0.588	0.280	0.525
VAD: arousal	Self	Gemma	48	1.000	1	0.000	0.525	0.232	0.485
VAD: arousal	Self	Granite	7	0.175	10	1.000	0.894	0.766	0.881
VAD: arousal	Self	Nemotron	9	0.161	8	0.750	0.614	0.153	0.609
VAD: arousal	Self	Qwen	5	0.156	1	0.000	0.437	0.101	0.462
VAD: dominance	Mixed	Gemma	40	0.833	1	0.000	0.873	0.753	0.854
VAD: dominance	Mixed	Granite	20	0.500	10	1.000	0.768	0.527	0.698
VAD: dominance	Mixed	Nemotron	37	0.661	5	0.500	0.782	0.601	0.769
VAD: dominance	Mixed	Qwen	5	0.156	1	0.000	0.710	0.398	0.719
VAD: dominance	Self	Gemma	32	0.667	1	0.000	0.804	0.580	0.793
VAD: dominance	Self	Granite	20	0.500	1	0.000	0.848	0.697	0.829
VAD: dominance	Self	Nemotron	9	0.161	1	0.000	0.662	0.406	0.695
VAD: dominance	Self	Qwen	5	0.156	8	0.750	0.697	0.444	0.666
Personality: openness	Mixed	Gemma	24	0.500	1	0.000	0.672	0.413	0.660
Personality: openness	Mixed	Granite	33	0.825	1	0.000	0.712	0.491	0.706
Personality: openness	Mixed	Nemotron	37	0.661	1	0.000	0.600	0.302	0.601
Personality: openness	Mixed	Qwen	11	0.344	1	0.000	0.621	0.317	0.583
Personality: openness	Self	Gemma	48	1.000	1	0.000	0.679	0.417	0.682
Personality: openness	Self	Granite	33	0.825	10	1.000	0.562	0.209	0.594
Personality: openness	Self	Nemotron	28	0.500	1	0.000	0.672	0.286	0.646
Personality: openness	Self	Qwen	16	0.500	1	0.000	0.602	0.272	0.592
Personality: conscientiousness	Mixed	Gemma	24	0.500	1	0.000	0.518	0.155	0.525
Personality: conscientiousness	Mixed	Granite	7	0.175	5	0.500	0.359	-0.026	0.342
Personality: conscientiousness	Mixed	Nemotron	19	0.339	5	0.500	0.423	0.006	0.395
Personality: conscientiousness	Mixed	Qwen	16	0.500	3	0.250	0.484	0.152	0.458
Personality: conscientiousness	Self	Gemma	32	0.667	1	0.000	0.335	-0.204	0.263
Personality: conscientiousness	Self	Granite	20	0.500	8	0.750	0.699	0.439	0.659
Personality: conscientiousness	Self	Nemotron	9	0.161	10	1.000	0.420	0.080	0.344
Personality: conscientiousness	Self	Qwen	11	0.344	1	0.000	0.376	0.013	0.344
Personality: extraversion	Mixed	Gemma	40	0.833	1	0.000	0.712	0.473	0.599
Personality: extraversion	Mixed	Granite	13	0.325	10	1.000	0.674	0.408	0.635
Personality: extraversion	Mixed	Nemotron	9	0.161	10	1.000	0.567	0.245	0.516
Personality: extraversion	Mixed	Qwen	11	0.344	1	0.000	0.642	0.320	0.562
Personality: extraversion	Self	Gemma	16	0.333	1	0.000	0.599	0.316	0.610
Personality: extraversion	Self	Granite	7	0.175	8	0.750	0.765	0.466	0.713
Personality: extraversion	Self	Nemotron	37	0.661	5	0.500	0.485	0.194	0.481
Personality: extraversion	Self	Qwen	11	0.344	1	0.000	0.507	0.144	0.392
Personality: agreeableness	Mixed	Gemma	32	0.667	10	1.000	0.751	0.521	0.737
Personality: agreeableness	Mixed	Granite	20	0.500	5	0.500	0.575	0.293	0.576
Personality: agreeableness	Mixed	Nemotron	28	0.500	5	0.500	0.725	0.505	0.672
Personality: agreeableness	Mixed	Qwen	16	0.500	10	1.000	0.699	0.425	0.628
Personality: agreeableness	Self	Gemma	24	0.500	1	0.000	0.749	0.548	0.745
Personality: agreeableness	Self	Granite	33	0.825	3	0.250	0.662	0.345	0.617
Personality: agreeableness	Self	Nemotron	37	0.661	1	0.000	0.679	0.382	0.682
Personality: agreeableness	Self	Qwen	11	0.344	10	1.000	0.694	0.443	0.674
Personality: neuroticism	Mixed	Gemma	32	0.667	1	0.000	0.657	0.242	0.441
Personality: neuroticism	Mixed	Granite	27	0.675	8	0.750	0.607	0.191	0.276
Personality: neuroticism	Mixed	Nemotron	37	0.661	1	0.000	0.589	0.301	0.480
Personality: neuroticism	Mixed	Qwen	32	1.000	10	1.000	0.317	-0.099	0.225
Personality: neuroticism	Self	Gemma	24	0.500	3	0.250	0.584	0.238	0.484
Personality: neuroticism	Self	Granite	20	0.500	10	1.000	0.753	0.169	0.331
Personality: neuroticism	Self	Nemotron	28	0.500	1	0.000	0.370	-0.054	0.395
Personality: neuroticism	Self	Qwen	32	1.000	10	1.000	0.516	0.223	0.355

Continued on next page

Table 3 continued

Target	Condition	Model	Layer	Depth	Turn	Progress	Pearson r	R^2	Spearman ρ
Semantic PC space	Mixed	Gemma	16	0.333	1	0.000	0.566	0.275	0.525
Semantic PC space	Mixed	Granite	13	0.325	8	0.750	0.572	0.316	0.552
Semantic PC space	Mixed	Nemotron	9	0.161	1	0.000	0.522	0.245	0.495
Semantic PC space	Mixed	Qwen	5	0.156	1	0.000	0.557	0.288	0.538
Semantic PC space	Self	Gemma	40	0.833	1	0.000	0.494	0.108	0.458
Semantic PC space	Self	Granite	13	0.325	1	0.000	0.601	0.332	0.574
Semantic PC space	Self	Nemotron	9	0.161	5	0.500	0.442	0.139	0.436
Semantic PC space	Self	Qwen	5	0.156	1	0.000	0.446	0.126	0.432
Style PC space	Mixed	Gemma	8	0.167	10	1.000	0.557	0.282	0.532
Style PC space	Mixed	Granite	7	0.175	10	1.000	0.568	0.290	0.539
Style PC space	Mixed	Nemotron	9	0.161	10	1.000	0.365	0.080	0.339
Style PC space	Mixed	Qwen	5	0.156	1	0.000	0.475	0.185	0.452
Style PC space	Self	Gemma	8	0.167	1	0.000	0.447	0.110	0.418
Style PC space	Self	Granite	7	0.175	10	1.000	0.555	0.279	0.535
Style PC space	Self	Nemotron	9	0.161	10	1.000	0.374	0.026	0.346
Style PC space	Self	Qwen	5	0.156	10	1.000	0.416	0.094	0.396

E.2 Layer and conversation-time localization

To summarize when a representation was most separable, we first averaged self-play and mixed-play probe scores and macro-averaged the five personality traits and three VAD dimensions. For each displayed variable and model, we then selected the single layer–checkpoint cell with the largest Pearson correlation and held that layer fixed over the trajectory. The benchmark was $\max(0.10, 0.80 r_{\max})$. Entry is the first checkpoint in the contiguous above-benchmark run containing the peak. Exit is the first later checkpoint below the benchmark, or the final observed checkpoint when no later crossing occurred. These windows are descriptive and are not significance tests or evidence that a representation was literally created or destroyed.

E.3 Self-play trajectory inertia

Trajectory inertia was estimated in cross-fitted probe-output coordinates by a ridge autoregression of the current state on the speaker's preceding state and the partner's preceding state. The inertia coefficient is the mean diagonal of the speaker-history coefficient block; β_{partner} is the corresponding mean diagonal for partner history. The stored RQ1 persistence analysis contains self-play trajectories only. Table 5 therefore reports descriptive means and ranges over the six sampled layers and does not imply a self-play–mixed-play comparison. Each layer-specific regression used 1,080 within-speaker transitions.

Table 5: Self-play trajectory persistence across the six sampled activation depths.

Target	Model	Mean inertia	Layer min.	Layer max.	Mean β_{partner}	Transitions/layer
Stance (subjective)	Gemma	0.651	0.528	0.736	-0.235	1080
Stance (subjective)	Granite	0.580	0.412	0.705	0.037	1080
Stance (subjective)	Nemotron	0.644	0.495	0.702	0.100	1080
Stance (subjective)	Qwen	0.647	0.528	0.692	-0.221	1080
Stance (objective)	Gemma	0.672	0.514	0.782	-0.210	1080
Stance (objective)	Granite	0.645	0.490	0.748	0.021	1080
Stance (objective)	Nemotron	0.593	0.474	0.636	0.212	1080
Stance (objective)	Qwen	0.682	0.551	0.740	-0.170	1080
VAD: valence	Gemma	0.389	0.290	0.484	-0.066	1080
VAD: valence	Granite	0.424	0.400	0.446	0.282	1080
VAD: valence	Nemotron	0.239	0.211	0.250	0.333	1080
VAD: valence	Qwen	0.318	0.256	0.374	0.030	1080
VAD: arousal	Gemma	0.092	0.030	0.150	0.112	1080
VAD: arousal	Granite	0.336	0.306	0.362	0.380	1080
VAD: arousal	Nemotron	0.198	0.155	0.221	0.205	1080
VAD: arousal	Qwen	0.084	0.049	0.111	0.179	1080
VAD: dominance	Gemma	0.177	0.129	0.227	-0.048	1080
VAD: dominance	Granite	0.224	0.207	0.238	0.157	1080
VAD: dominance	Nemotron	0.141	0.115	0.161	0.184	1080
VAD: dominance	Qwen	0.184	0.132	0.218	0.035	1080
Personality: openness	Gemma	0.376	0.202	0.484	0.033	1080
Personality: openness	Granite	0.318	0.209	0.385	0.303	1080
Personality: openness	Nemotron	0.277	0.240	0.303	0.430	1080

Continued on next page

Table 5 continued

Target	Model	Mean inertia	Layer min.	Layer max.	Mean β_{partner}	Transitions/layer
Personality: openness	Qwen	0.308	0.229	0.373	0.161	1080
Personality: conscientiousness	Gemma	0.180	0.065	0.223	0.092	1080
Personality: conscientiousness	Granite	0.249	0.216	0.297	0.344	1080
Personality: conscientiousness	Nemotron	0.206	0.143	0.272	0.387	1080
Personality: conscientiousness	Qwen	0.164	0.140	0.204	0.204	1080
Personality: extraversion	Gemma	0.336	0.241	0.381	-0.009	1080
Personality: extraversion	Granite	0.379	0.357	0.396	0.400	1080
Personality: extraversion	Nemotron	0.282	0.253	0.301	0.372	1080
Personality: extraversion	Qwen	0.336	0.235	0.383	0.098	1080
Personality: agreeableness	Gemma	0.471	0.321	0.545	-0.152	1080
Personality: agreeableness	Granite	0.433	0.366	0.487	0.199	1080
Personality: agreeableness	Nemotron	0.337	0.315	0.352	0.219	1080
Personality: agreeableness	Qwen	0.537	0.417	0.587	-0.135	1080
Personality: neuroticism	Gemma	0.359	0.196	0.465	0.048	1080
Personality: neuroticism	Granite	0.360	0.284	0.402	0.143	1080
Personality: neuroticism	Nemotron	0.327	0.251	0.360	0.252	1080
Personality: neuroticism	Qwen	0.343	0.298	0.390	0.102	1080
Semantic PC space	Gemma	0.058	0.022	0.072	0.155	1080
Semantic PC space	Granite	0.131	0.123	0.138	0.282	1080
Semantic PC space	Nemotron	0.145	0.140	0.151	0.285	1080
Semantic PC space	Qwen	0.075	0.069	0.084	0.219	1080
Style PC space	Gemma	0.139	0.096	0.181	0.121	1080
Style PC space	Granite	0.235	0.205	0.246	0.252	1080
Style PC space	Nemotron	0.195	0.181	0.218	0.294	1080
Style PC space	Qwen	0.126	0.107	0.162	0.181	1080

E.4 DIRECTIONAL PARTNERWARD PULL

For mixed-play state $x_{A|B}$ and matched self-play references $x_{A|A}$ and $x_{B|B}$, partnerward pull is the scalar projection $\alpha = ((x_{A|B} - x_{A|A})^\top (x_{B|B} - x_{A|A}))/\|x_{B|B} - x_{A|A}\|^2$. Values of zero indicate no movement along the matched-self axis, values between zero and one indicate movement toward the partner reference, negative values indicate movement away, and values above one overshoot the partner reference.

The RQ1 directional artifacts do not retain the reference distance in the denominator. Because nearly coincident references can produce extreme α, Table 6 treats the median as the primary descriptive estimate and reports the arithmetic mean and tail diagnostics as sensitivity information. No denominator-based observations were removed. Each directed dyad contains 1,800 observations: 20 topics $\times$ 3 seeds $\times$ 5 checkpoints $\times$ 6 depths. Inferential confidence intervals are not available in the stored artifacts.

Table 6: Directional partnerward pull by target and ordered model dyad. Medians are the primary summaries because the projection ratio is heavy-tailed.

Target	Speaker	Partner	Median α	Mean α	5th pct.	95th pct.	$\lvert\alpha\rvert > 10$ (%)	n
Stance (subjective)	Gemma	Granite	-0.006	2.618	-2.284	2.880	2.4	1800
Stance (subjective)	Gemma	Nemotron	-0.024	-0.306	-1.396	0.621	1.7	1800
Stance (subjective)	Gemma	Qwen	0.021	0.047	-1.716	0.698	1.2	1800
Stance (subjective)	Granite	Gemma	0.157	1.326	-1.658	1.933	1.6	1800
Stance (subjective)	Granite	Nemotron	0.388	0.617	-3.209	4.521	3.9	1800
Stance (subjective)	Granite	Qwen	0.292	0.359	-4.124	4.023	4.1	1800
Stance (subjective)	Nemotron	Gemma	0.140	0.096	-1.166	1.233	1.3	1800
Stance (subjective)	Nemotron	Granite	0.478	0.126	-4.843	5.505	4.9	1800
Stance (subjective)	Nemotron	Qwen	0.361	0.844	-1.969	2.937	2.3	1800
Stance (subjective)	Qwen	Gemma	-0.082	-0.743	-1.504	0.449	1.2	1800
Stance (subjective)	Qwen	Granite	0.056	0.227	-4.671	4.976	4.9	1800
Stance (subjective)	Qwen	Nemotron	-0.093	-0.737	-2.312	1.899	2.0	1800
Stance (objective)	Gemma	Granite	-0.033	-0.255	-2.647	1.627	2.4	1800
Stance (objective)	Gemma	Nemotron	-0.017	-0.055	-2.248	1.257	1.8	1800
Stance (objective)	Gemma	Qwen	-0.006	-0.279	-1.328	0.564	0.7	1800
Stance (objective)	Granite	Gemma	0.187	1.317	-1.670	2.527	2.1	1800
Stance (objective)	Granite	Nemotron	0.329	-0.003	-3.392	3.884	3.4	1800
Stance (objective)	Granite	Qwen	0.234	0.466	-2.823	3.481	3.2	1800

Continued on next page

Table 6 continued

Target	Speaker	Partner	Median α	Mean α	5th pct.	95th pct.	$\lvert\alpha\rvert > 10$ (%)	n
Stance (objective)	Nemotron	Gemma	0.265	0.043	-1.091	1.485	1.3	1800
Stance (objective)	Nemotron	Granite	0.106	-0.099	-3.276	3.591	3.4	1800
Stance (objective)	Nemotron	Qwen	0.341	-0.036	-1.509	2.259	2.1	1800
Stance (objective)	Qwen	Gemma	-0.053	0.031	-1.112	0.407	0.6	1800
Stance (objective)	Qwen	Granite	-0.063	-1.986	-2.926	2.541	3.6	1800
Stance (objective)	Qwen	Nemotron	-0.078	0.540	-2.998	2.626	3.0	1800
VAD: valence	Gemma	Granite	0.450	0.372	-2.345	2.729	2.7	1800
VAD: valence	Gemma	Nemotron	0.581	1.315	-3.309	4.419	4.1	1800
VAD: valence	Gemma	Qwen	0.240	-9.112	-4.435	4.496	4.4	1800
VAD: valence	Granite	Gemma	0.316	0.897	-2.433	3.093	2.9	1800
VAD: valence	Granite	Nemotron	0.422	-0.388	-3.638	4.350	3.4	1800
VAD: valence	Granite	Qwen	0.305	6.398	-2.381	3.174	2.8	1800
VAD: valence	Nemotron	Gemma	0.151	-0.895	-4.779	4.819	4.7	1800
VAD: valence	Nemotron	Granite	0.331	2.181	-3.103	4.132	3.3	1800
VAD: valence	Nemotron	Qwen	0.211	-0.435	-5.091	4.240	5.2	1800
VAD: valence	Qwen	Gemma	0.271	-37.332	-4.838	5.423	4.9	1800
VAD: valence	Qwen	Granite	0.417	21.027	-2.823	3.371	3.1	1800
VAD: valence	Qwen	Nemotron	0.643	1.025	-4.180	5.411	4.8	1800
VAD: arousal	Gemma	Granite	0.341	0.358	-4.017	3.882	4.2	1800
VAD: arousal	Gemma	Nemotron	0.533	-0.665	-5.080	5.414	5.3	1800
VAD: arousal	Gemma	Qwen	0.434	-0.027	-5.558	5.493	5.7	1800
VAD: arousal	Granite	Gemma	0.475	0.140	-4.603	3.797	4.4	1800
VAD: arousal	Granite	Nemotron	0.448	0.567	-4.464	5.481	5.1	1800
VAD: arousal	Granite	Qwen	0.547	1.459	-3.035	4.700	4.1	1800
VAD: arousal	Nemotron	Gemma	0.447	-0.110	-5.290	6.079	5.9	1800
VAD: arousal	Nemotron	Granite	0.329	-0.471	-3.849	4.050	3.7	1800
VAD: arousal	Nemotron	Qwen	0.444	1.666	-5.037	5.569	5.3	1800
VAD: arousal	Qwen	Gemma	0.473	0.632	-4.657	5.163	5.2	1800
VAD: arousal	Qwen	Granite	0.389	-0.302	-3.876	4.393	3.8	1800
VAD: arousal	Qwen	Nemotron	0.567	0.516	-4.926	5.803	6.0	1800
VAD: dominance	Gemma	Granite	0.418	0.836	-2.632	3.796	3.6	1800
VAD: dominance	Gemma	Nemotron	0.586	0.197	-3.660	4.575	4.1	1800
VAD: dominance	Gemma	Qwen	0.191	-0.088	-3.821	3.858	3.8	1800
VAD: dominance	Granite	Gemma	0.364	-0.593	-2.961	3.237	3.4	1800
VAD: dominance	Granite	Nemotron	0.416	1.041	-3.811	3.757	3.4	1800
VAD: dominance	Granite	Qwen	0.341	2.097	-2.902	3.615	3.3	1800
VAD: dominance	Nemotron	Gemma	0.287	0.454	-4.655	5.826	5.1	1800
VAD: dominance	Nemotron	Granite	0.363	0.575	-3.039	4.129	3.5	1800
VAD: dominance	Nemotron	Qwen	0.340	-1.063	-5.891	6.720	6.4	1800
VAD: dominance	Qwen	Gemma	0.272	0.223	-4.385	4.723	3.9	1800
VAD: dominance	Qwen	Granite	0.443	-1.954	-3.646	3.945	3.1	1800
VAD: dominance	Qwen	Nemotron	0.619	0.120	-4.943	6.263	5.6	1800
Personality: openness	Gemma	Granite	0.360	-2.350	-3.662	4.863	4.1	1800
Personality: openness	Gemma	Nemotron	0.341	-0.080	-2.901	3.214	2.4	1800
Personality: openness	Gemma	Qwen	0.416	0.444	-5.067	5.625	4.4	1800
Personality: openness	Granite	Gemma	0.375	-1.443	-4.668	4.113	4.1	1800
Personality: openness	Granite	Nemotron	0.480	-12.221	-5.086	4.503	4.5	1800
Personality: openness	Granite	Qwen	0.500	-0.257	-3.968	5.230	4.7	1800
Personality: openness	Nemotron	Gemma	0.329	0.320	-2.712	3.525	2.7	1800
Personality: openness	Nemotron	Granite	0.313	3.852	-4.493	4.568	5.4	1800
Personality: openness	Nemotron	Qwen	0.438	-0.103	-3.102	3.786	3.1	1800
Personality: openness	Qwen	Gemma	0.225	-0.558	-5.198	4.810	5.6	1800
Personality: openness	Qwen	Granite	0.276	1.090	-3.947	5.159	4.7	1800
Personality: openness	Qwen	Nemotron	0.262	0.331	-3.743	2.967	2.9	1800
Personality: conscientiousness	Gemma	Granite	0.279	1.204	-5.358	5.619	5.9	1800
Personality: conscientiousness	Gemma	Nemotron	0.415	0.553	-5.294	6.363	5.8	1800
Personality: conscientiousness	Gemma	Qwen	0.432	138.487	-6.114	7.005	6.4	1800
Personality: conscientiousness	Granite	Gemma	0.434	-0.551	-4.186	5.139	4.8	1800
Personality: conscientiousness	Granite	Nemotron	0.648	1.566	-3.778	5.087	4.6	1800
Personality: conscientiousness	Granite	Qwen	0.449	-0.335	-4.479	4.120	3.8	1800
Personality: conscientiousness	Nemotron	Gemma	0.332	-0.626	-4.802	4.889	4.9	1800
Personality: conscientiousness	Nemotron	Granite	0.331	-0.334	-3.997	5.406	4.3	1800
Personality: conscientiousness	Nemotron	Qwen	0.404	1.371	-3.832	4.908	4.8	1800
Personality: conscientiousness	Qwen	Gemma	0.488	55.465	-4.800	6.545	6.2	1800

Continued on next page

Table 6 continued

Target	Speaker	Partner	Median α	Mean α	5th pct.	95th pct.	$\lvert\alpha\rvert > 10$ (%)	n
Personality: conscientiousness	Qwen	Granite	0.292	-0.998	-4.417	5.032	4.6	1800
Personality: conscientiousness	Qwen	Nemotron	0.436	-1.590	-4.282	6.240	5.4	1800
Personality: extraversion	Gemma	Granite	0.236	1.212	-5.832	7.110	6.4	1800
Personality: extraversion	Gemma	Nemotron	0.377	0.096	-4.663	5.869	5.6	1800
Personality: extraversion	Gemma	Qwen	0.288	-0.230	-4.749	5.728	5.9	1800
Personality: extraversion	Granite	Gemma	0.539	0.168	-3.420	5.708	5.2	1800
Personality: extraversion	Granite	Nemotron	0.608	0.686	-3.073	5.513	4.5	1800
Personality: extraversion	Granite	Qwen	0.551	1.144	-3.473	4.238	3.6	1800
Personality: extraversion	Nemotron	Gemma	0.429	0.210	-6.616	6.172	6.1	1800
Personality: extraversion	Nemotron	Granite	0.291	0.908	-4.184	4.868	4.8	1800
Personality: extraversion	Nemotron	Qwen	0.487	2.241	-5.705	7.019	6.4	1800
Personality: extraversion	Qwen	Gemma	0.308	-0.507	-4.545	5.493	5.6	1800
Personality: extraversion	Qwen	Granite	0.271	-1.440	-5.687	4.647	5.0	1800
Personality: extraversion	Qwen	Nemotron	0.534	-0.008	-5.391	6.000	6.3	1800
Personality: agreeableness	Gemma	Granite	0.529	-0.820	-4.046	4.729	4.6	1800
Personality: agreeableness	Gemma	Nemotron	0.443	0.781	-2.698	3.870	3.4	1800
Personality: agreeableness	Gemma	Qwen	0.282	0.730	-2.405	2.564	2.8	1800
Personality: agreeableness	Granite	Gemma	0.272	1.204	-3.733	3.761	3.6	1800
Personality: agreeableness	Granite	Nemotron	0.564	0.112	-4.105	6.385	5.4	1800
Personality: agreeableness	Granite	Qwen	0.353	-0.559	-3.560	3.828	3.3	1800
Personality: agreeableness	Nemotron	Gemma	0.310	-0.010	-2.765	3.237	3.1	1800
Personality: agreeableness	Nemotron	Granite	0.391	-2.622	-5.638	4.549	5.4	1800
Personality: agreeableness	Nemotron	Qwen	0.326	0.951	-3.415	4.515	4.6	1800
Personality: agreeableness	Qwen	Gemma	0.059	0.690	-2.303	2.088	2.7	1800
Personality: agreeableness	Qwen	Granite	0.397	0.677	-3.993	5.045	3.7	1800
Personality: agreeableness	Qwen	Nemotron	0.571	0.108	-3.479	5.902	5.0	1800
Personality: neuroticism	Gemma	Granite	0.556	3.755	-4.717	5.632	5.2	1800
Personality: neuroticism	Gemma	Nemotron	0.688	0.822	-3.723	5.062	4.4	1800
Personality: neuroticism	Gemma	Qwen	0.336	0.927	-5.502	5.216	4.9	1800
Personality: neuroticism	Granite	Gemma	0.281	-18.891	-4.953	4.763	5.0	1800
Personality: neuroticism	Granite	Nemotron	0.699	0.176	-6.281	7.619	6.9	1800
Personality: neuroticism	Granite	Qwen	0.175	0.798	-4.383	4.740	3.8	1800
Personality: neuroticism	Nemotron	Gemma	0.150	0.066	-4.586	4.182	3.9	1800
Personality: neuroticism	Nemotron	Granite	0.241	1.355	-6.146	6.924	6.3	1800
Personality: neuroticism	Nemotron	Qwen	0.115	0.553	-4.537	4.119	4.6	1800
Personality: neuroticism	Qwen	Gemma	0.436	-0.201	-3.527	4.397	3.7	1800
Personality: neuroticism	Qwen	Granite	0.755	1.692	-2.418	4.003	3.0	1800
Personality: neuroticism	Qwen	Nemotron	0.879	0.217	-2.699	4.231	3.6	1800
Semantic PC space	Gemma	Granite	0.313	0.316	-0.135	0.777	0.0	1800
Semantic PC space	Gemma	Nemotron	0.334	0.350	-0.142	0.897	0.0	1800
Semantic PC space	Gemma	Qwen	0.392	0.399	-0.369	1.176	0.0	1800
Semantic PC space	Granite	Gemma	0.411	0.418	0.014	0.868	0.0	1800
Semantic PC space	Granite	Nemotron	0.455	0.453	-0.031	0.925	0.0	1800
Semantic PC space	Granite	Qwen	0.377	0.377	-0.012	0.766	0.0	1800
Semantic PC space	Nemotron	Gemma	0.449	0.448	-0.068	0.951	0.0	1800
Semantic PC space	Nemotron	Granite	0.339	0.348	-0.148	0.864	0.0	1800
Semantic PC space	Nemotron	Qwen	0.453	0.451	-0.068	0.962	0.0	1800
Semantic PC space	Qwen	Gemma	0.452	0.462	-0.199	1.144	0.0	1800
Semantic PC space	Qwen	Granite	0.265	0.262	-0.133	0.657	0.0	1800
Semantic PC space	Qwen	Nemotron	0.342	0.340	-0.130	0.818	0.0	1800
Style PC space	Gemma	Granite	0.253	0.258	-0.188	0.698	0.0	1800
Style PC space	Gemma	Nemotron	0.259	0.254	-0.153	0.663	0.0	1800
Style PC space	Gemma	Qwen	0.443	0.438	-0.297	1.144	0.0	1800
Style PC space	Granite	Gemma	0.245	0.265	-0.136	0.732	0.0	1800
Style PC space	Granite	Nemotron	0.519	0.508	-0.242	1.187	0.0	1800
Style PC space	Granite	Qwen	0.330	0.353	-0.112	0.890	0.0	1800
Style PC space	Nemotron	Gemma	0.211	0.228	-0.103	0.609	0.0	1800
Style PC space	Nemotron	Granite	0.364	0.374	-0.266	1.081	0.0	1800
Style PC space	Nemotron	Qwen	0.292	0.294	-0.158	0.775	0.0	1800
Style PC space	Qwen	Gemma	0.367	0.360	-0.353	1.042	0.0	1800
Style PC space	Qwen	Granite	0.284	0.273	-0.270	0.766	0.0	1800
Style PC space	Qwen	Nemotron	0.310	0.310	-0.183	0.813	0.0	1800

E.5 Complementary trajectory-geometry decomposition

Signed pull alone does not distinguish weak net accommodation from large but oppositely directed movements. Table 7 therefore also reports median absolute pull, pair contraction, absolute directional dominance, and normalized off-axis displacement. Pair contraction is $1 - \|x_{A|B} - x_{B|A}\| / \|x_{A|A} - x_{B|B}\|$; positive values indicate that the mixed-play pair is closer than its matched

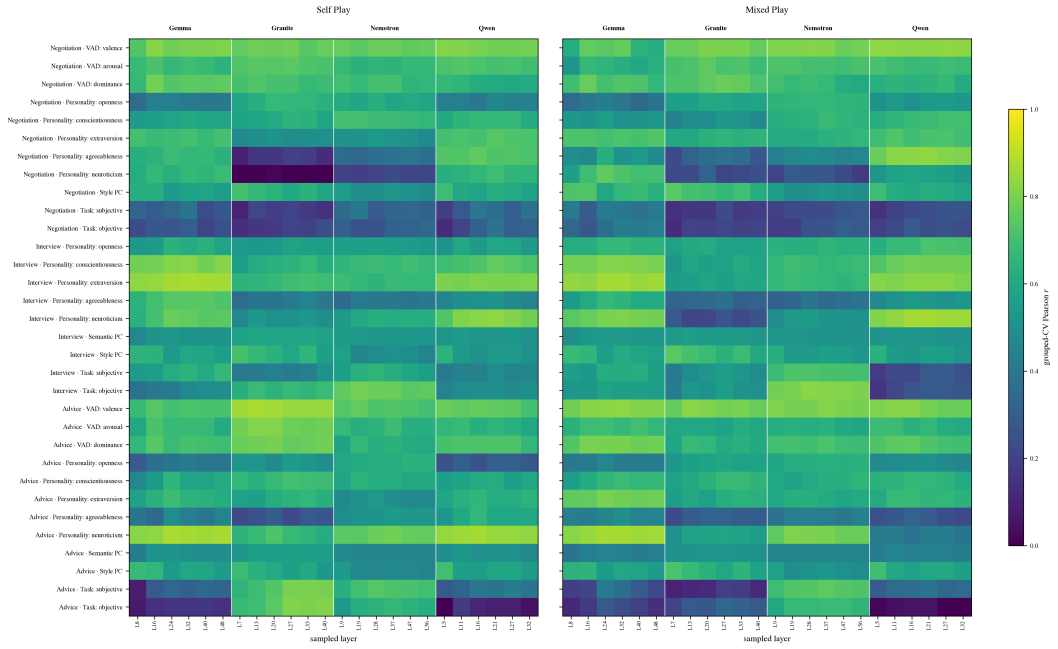


Figure 6: Held-out decoding quality in advice, interview, and negotiation. Each cell averages Pearson r over the five conversation checkpoints. Debate decoding quality is reported in Appendix E.

self-play references. Directional dominance is the reverse-direction pull minus the focal-direction pull, and its absolute value measures asymmetry without arbitrarily privileging a direction. Off-axis displacement is reported only for the multivariate semantic and style PC spaces; in a scalar decoded space, an orthogonal component is zero by construction. All entries are descriptive medians over topics, seeds, checkpoints, depths, and applicable model dyads.

F RQ2: Attunement Across Four Conversational Settings

RQ2 compares debate, advice, interview, and negotiation, with 600 conversations and 12,000 model responses per setting. The debate measurements are the same decoded-state geometry analyzed in RQ1; Figure 4 reuses them only to make the cross-setting comparison. To avoid duplicating Appendix E, the detailed tables below report only the three additional settings—advice, interview, and negotiation. Across all settings, decoded states are out-of-fold predictions from fivefold cross-validation grouped by conversation, and geometry is evaluated at six normalized model depths.

F.1 Decoding quality in the additional settings

Figure 6 reports held-out Pearson correlation for every decoded variable, model, and layer in advice, interview, and negotiation, separately for self-play and mixed-play. Table 8 gives a compact variable-level summary. The peak entries are descriptive maxima over condition, model, and layer, rather than estimates from a second confirmatory holdout.

Table 8: Compact decoding-quality summary for the three additional settings by decoded variable.

Setting	Variable	Mean r	Mean R^2	Peak r	Peak R^2	Condition	Model	Layer
Negotiation	VAD: valence	0.777	0.570	0.833	0.688	mixed play	Qwen	27
Negotiation	VAD: arousal	0.689	0.429	0.752	0.539	mixed play	Granite	20
Negotiation	VAD: dominance	0.688	0.384	0.791	0.531	self play	Gemma	16
Negotiation	Personality: openness	0.529	0.151	0.671	0.365	self play	Granite	20
Negotiation	Personality: conscientiousness	0.607	0.275	0.704	0.463	self play	Nemotron	9
Negotiation	Personality: extraversion	0.639	0.342	0.740	0.515	self play	Qwen	21

Continued on next page

Table 8 continued

Setting	Variable	Mean r	Mean R^2	Peak r	Peak R^2	Condition	Model	Layer
Negotiation	Personality: agreeableness	0.501	0.091	0.832	0.673	mixed play	Qwen	21
Negotiation	Personality: neuroticism	0.403	-0.075	0.756	0.483	mixed play	Gemma	24
Negotiation	Style PC	0.606	0.384	0.743	0.567	mixed play	Granite	7
Negotiation	Task: subjective	0.268	-0.242	0.409	-0.028	mixed play	Gemma	48
Negotiation	Task: objective	0.256	-0.263	0.433	0.009	mixed play	Gemma	48
Interview	Personality: openness	0.603	0.273	0.715	0.455	mixed play	Qwen	21
Interview	Personality: conscientiousness	0.707	0.440	0.822	0.656	self play	Gemma	32
Interview	Personality: extraversion	0.739	0.504	0.873	0.756	self play	Gemma	32
Interview	Personality: agreeableness	0.475	0.040	0.741	0.533	self play	Gemma	32
Interview	Personality: neuroticism	0.629	0.256	0.870	0.744	mixed play	Qwen	21
Interview	Semantic PC	0.533	0.289	0.592	0.351	self play	Granite	13
Interview	Style PC	0.586	0.360	0.744	0.561	mixed play	Granite	7
Interview	Task: subjective	0.516	0.127	0.722	0.488	mixed play	Nemotron	19
Interview	Task: objective	0.557	0.179	0.826	0.674	mixed play	Nemotron	28
Advice	VAD: valence	0.780	0.579	0.872	0.756	self play	Granite	13
Advice	VAD: arousal	0.660	0.372	0.815	0.655	self play	Granite	13
Advice	VAD: dominance	0.705	0.420	0.800	0.601	mixed play	Gemma	16
Advice	Personality: openness	0.486	0.063	0.641	0.344	self play	Nemotron	47
Advice	Personality: conscientiousness	0.625	0.305	0.701	0.451	self play	Granite	27
Advice	Personality: extraversion	0.628	0.307	0.798	0.608	mixed play	Gemma	24
Advice	Personality: agreeableness	0.401	-0.274	0.662	0.373	self play	Qwen	16
Advice	Personality: neuroticism	0.717	0.433	0.874	0.753	self play	Gemma	24
Advice	Semantic PC	0.480	0.230	0.570	0.340	self play	Granite	27
Advice	Style PC	0.580	0.351	0.715	0.545	mixed play	Granite	7
Advice	Task: subjective	0.446	0.042	0.800	0.621	self play	Granite	27
Advice	Task: objective	0.344	-0.141	0.812	0.645	self play	Granite	27

F.2 Detailed inertia and partnerward pull in the additional settings

Table 9 disaggregates the advice, interview, and negotiation geometry by setting, variable family, and speaker model. Inertia is averaged over decoded dimensions, conditions, and the six layers; its reported range is the minimum and maximum stored cell estimate. Pull excludes observations whose matched self-play reference distance is below 10^{-4}. Pull intervals are normal-approximation intervals over conversation-level mean pull and should be read as descriptive uncertainty, not confirmatory tests.

Table 9: Model-level decoded-state inertia and denominator-filtered partnerward pull in the three additional settings.

Setting	Family	Model	Inertia	Min.	Max.	Mean pull	Median pull	95% interval	Convs.	Excl.
Negotiation	VAD	Gemma	0.295	0.259	0.355	0.499	0.432	[0.214, 0.784]	180	19
Negotiation	VAD	Granite	0.226	0.079	0.410	0.279	0.177	[-0.043, 0.601]	180	26
Negotiation	VAD	Nemotron	0.252	0.182	0.383	0.590	0.488	[0.346, 0.833]	180	20
Negotiation	VAD	Qwen	0.385	0.187	0.577	0.368	0.475	[0.081, 0.654]	180	19
Negotiation	Personality	Gemma	0.510	0.361	0.677	0.000	0.330	[-0.762, 0.763]	180	3
Negotiation	Personality	Granite	0.294	0.010	0.551	0.009	0.316	[-0.507, 0.526]	180	7
Negotiation	Personality	Nemotron	0.329	0.074	0.598	0.300	0.451	[-0.528, 1.128]	180	4
Negotiation	Personality	Qwen	0.606	0.327	0.751	0.428	0.473	[-0.337, 1.192]	180	4
Negotiation	Style	Gemma	0.211	0.170	0.260	0.222	0.208	[0.200, 0.245]	180	0
Negotiation	Style	Granite	0.192	0.143	0.256	0.267	0.258	[0.248, 0.286]	180	0
Negotiation	Style	Nemotron	0.210	0.146	0.281	0.508	0.494	[0.488, 0.527]	180	0
Negotiation	Style	Qwen	0.317	0.264	0.375	0.254	0.242	[0.233, 0.274]	180	0
Negotiation	Task	Gemma	0.335	0.240	0.399	0.733	0.008	[-0.188, 1.655]	180	2
Negotiation	Task	Granite	0.278	0.147	0.457	-0.471	0.168	[-1.659, 0.716]	180	1
Negotiation	Task	Nemotron	0.325	0.203	0.477	-0.102	-0.121	[-0.744, 0.540]	180	0
Negotiation	Task	Qwen	0.472	0.305	0.602	0.512	0.246	[0.035, 0.989]	180	1
Interview	Personality	Gemma	0.629	0.408	0.830	0.348	0.401	[-0.170, 0.865]	180	3
Interview	Personality	Granite	0.318	0.178	0.514	0.048	0.342	[-0.583, 0.679]	180	3
Interview	Personality	Nemotron	0.405	0.200	0.606	0.717	0.433	[-0.027, 1.462]	180	6
Interview	Personality	Qwen	0.560	0.244	0.786	-0.020	0.284	[-0.863, 0.824]	180	2
Interview	Semantic	Gemma	0.124	0.071	0.174	0.412	0.421	[0.387, 0.436]	180	0
Interview	Semantic	Granite	0.185	0.141	0.240	0.313	0.305	[0.292, 0.333]	180	0
Interview	Semantic	Nemotron	0.170	0.130	0.214	0.403	0.398	[0.381, 0.425]	180	0
Interview	Semantic	Qwen	0.112	0.098	0.127	0.392	0.398	[0.367, 0.417]	180	0
Interview	Style	Gemma	0.256	0.212	0.294	0.315	0.304	[0.296, 0.334]	180	0
Interview	Style	Granite	0.236	0.202	0.263	0.282	0.259	[0.255, 0.309]	180	0
Interview	Style	Nemotron	0.226	0.162	0.297	0.336	0.313	[0.314, 0.358]	180	0
Interview	Style	Qwen	0.194	0.157	0.232	0.298	0.296	[0.279, 0.316]	180	0

Continued on next page

Table 9 continued

Setting	Family	Model	Inertia	Min.	Max.	Mean pull	Median pull	95% interval	Convs.	Excl.
Interview	Task	Gemma	0.564	0.418	0.728	0.531	0.066	[-0.101, 1.163]	180	0
Interview	Task	Granite	0.574	0.401	0.688	0.084	-0.046	[-0.106, 0.274]	180	0
Interview	Task	Nemotron	0.648	0.552	0.728	-0.116	-0.075	[-0.331, 0.099]	180	0
Interview	Task	Qwen	0.468	0.237	0.712	-0.082	0.041	[-0.433, 0.269]	180	0
Advice	VAD	Gemma	0.519	0.419	0.622	0.368	0.349	[0.011, 0.726]	180	11
Advice	VAD	Granite	0.338	0.228	0.533	0.112	0.145	[-0.158, 0.382]	180	11
Advice	VAD	Nemotron	0.464	0.296	0.542	0.350	0.338	[0.002, 0.699]	180	12
Advice	VAD	Qwen	0.424	0.272	0.549	0.857	0.602	[0.501, 1.213]	180	8
Advice	Personality	Gemma	0.561	0.322	0.851	1.091	0.216	[-0.163, 2.346]	180	0
Advice	Personality	Granite	0.379	0.127	0.583	0.231	0.445	[-0.292, 0.754]	180	3
Advice	Personality	Nemotron	0.498	0.318	0.732	-0.137	0.350	[-1.185, 0.911]	180	2
Advice	Personality	Qwen	0.465	0.234	0.811	0.406	0.599	[-0.296, 1.107]	180	1
Advice	Semantic	Gemma	0.166	0.112	0.218	0.373	0.378	[0.354, 0.392]	180	0
Advice	Semantic	Granite	0.212	0.177	0.261	0.369	0.357	[0.347, 0.390]	180	0
Advice	Semantic	Nemotron	0.215	0.161	0.264	0.401	0.387	[0.376, 0.425]	180	0
Advice	Semantic	Qwen	0.160	0.113	0.204	0.437	0.432	[0.415, 0.459]	180	0
Advice	Style	Gemma	0.203	0.146	0.270	0.284	0.281	[0.267, 0.302]	180	0
Advice	Style	Granite	0.254	0.217	0.285	0.307	0.285	[0.286, 0.327]	180	0
Advice	Style	Nemotron	0.310	0.255	0.364	0.394	0.380	[0.372, 0.416]	180	0
Advice	Style	Qwen	0.195	0.149	0.240	0.319	0.314	[0.301, 0.338]	180	0
Advice	Task	Gemma	0.412	0.268	0.598	0.099	-0.024	[-0.264, 0.461]	180	1
Advice	Task	Granite	0.548	0.214	0.760	0.257	-0.090	[-0.890, 1.404]	180	0
Advice	Task	Nemotron	0.593	0.391	0.696	-0.065	-0.217	[-0.445, 0.316]	180	0
Advice	Task	Qwen	0.390	0.154	0.523	0.118	-0.025	[-0.234, 0.471]	180	1

F.3 Pull denominator sensitivity

Partnerward pull divides by the squared distance between the two matched self-play references. A very small denominator can therefore dominate an arithmetic mean. Table 10 reports the same balanced macro-average under several minimum-distance thresholds, together with observation-level robust summaries. Only 170 of 691,200 observations (0.025%) fall below 10^{-4}, yet their removal materially changes several family means; filtered rather than raw means are therefore used in the remaining detailed tables for the three additional settings.

Table 10: Sensitivity of macro partnerward pull to minimum reference distance.

Setting	Family	Raw	$\epsilon = 10^{-5}$	10^{-4}	10^{-3}	Median	10% trimmed	Rows	Excl.	Sign stable
Negotiation	VAD	9.342	0.290	0.433	0.405	0.388	0.402	64800	84	yes
Negotiation	Personality	-0.562	-0.562	0.184	0.438	0.325	0.367	108000	18	no
Negotiation	Style	0.313	0.313	0.313	0.313	0.287	0.302	21600	0	yes
Negotiation	Task	0.688	0.688	0.168	0.154	0.129	0.121	43200	4	yes
Interview	Personality	0.505	0.505	0.274	0.295	0.318	0.344	108000	14	yes
Interview	Semantic	0.380	0.380	0.380	0.380	0.359	0.373	21600	0	yes
Interview	Style	0.308	0.308	0.308	0.308	0.279	0.295	21600	0	yes
Interview	Task	0.104	0.104	0.104	0.058	0.028	-0.001	43200	0	yes
Advice	VAD	0.406	0.406	0.421	0.392	0.361	0.386	64800	42	yes
Advice	Personality	0.297	0.297	0.398	0.416	0.341	0.378	108000	6	yes
Advice	Semantic	0.395	0.395	0.395	0.395	0.388	0.390	21600	0	yes
Advice	Style	0.326	0.326	0.326	0.326	0.312	0.320	21600	0	yes
Advice	Task	0.908	0.908	0.102	-0.119	0.002	-0.038	43200	2	yes

F.4 Dimension-level dyadic geometry

Figure 7 retains individual decoded dimensions and ordered model dyads. Signed pull distinguishes accommodation from movement away; absolute pull measures the magnitude of axial movement; and contraction measures whether the two mixed-play states become closer than their matched self-play references. The figure shows substantial dimension-, setting-, and direction-specific heterogeneity that is hidden by family averages.

F.5 Condition, time, and role heterogeneity

For advice, interview, and negotiation, the following compact tables separate three sources of heterogeneity. Table 11 compares self- and mixed-play inertia; Table 12 shows when filtered partnerward pull appears; and Table 13 separates the two role directions in each setting.

Variable	Model	Selected depth	Entry	Exit	Peak r	Exit rule
Personality	Gemma	0.500	0.00	0.25	0.572	Below benchmark
Personality	Granite	0.833	0.00	1.00	0.483	Final checkpoint
Personality	Nemotron	0.500	0.00	0.25	0.456	Below benchmark
Personality	Qwen	0.500	0.00	0.25	0.475	Below benchmark
Semantic embeddings PC space	Gemma	0.333	0.00	0.25	0.530	Below benchmark
Semantic embeddings PC space	Granite	0.333	0.00	0.25	0.575	Below benchmark
Semantic embeddings PC space	Nemotron	0.167	0.00	1.00	0.452	Final checkpoint
Semantic embeddings PC space	Qwen	0.167	0.00	0.25	0.502	Below benchmark
Stance (subjective)	Gemma	0.667	0.00	1.00	0.956	Final checkpoint
Stance (subjective)	Granite	0.667	0.00	0.50	0.853	Below benchmark
Stance (subjective)	Nemotron	0.500	0.00	0.50	0.929	Below benchmark
Stance (subjective)	Qwen	0.333	0.00	1.00	0.840	Final checkpoint
Style PC space	Gemma	0.167	0.00	1.00	0.496	Final checkpoint
Style PC space	Granite	0.167	0.50	1.00	0.561	Final checkpoint
Style PC space	Nemotron	0.167	0.50	1.00	0.369	Final checkpoint
Style PC space	Qwen	0.167	0.00	0.50	0.433	Below benchmark
VAD	Gemma	0.333	0.00	0.25	0.758	Below benchmark
VAD	Granite	0.167	0.50	1.00	0.827	Final checkpoint
VAD	Nemotron	0.167	0.00	1.00	0.656	Final checkpoint
VAD	Qwen	0.167	0.00	0.25	0.618	Below benchmark

Table 4: Fixed-layer temporal localization for the five paper-facing RQ1 variable families. Depth is normalized by each model's total number of transformer blocks.

Target	Signed pull	Absolute pull	Pair contraction	|dominance|	Off-axis
Stance (subjective)	0.083	0.421	0.015	0.633	–
Stance (objective)	0.059	0.376	0.036	0.568	–
VAD: valence	0.355	0.742	0.222	1.067	–
VAD: arousal	0.452	0.871	0.185	1.283	–
VAD: dominance	0.379	0.789	0.192	1.122	–
Personality: openness	0.354	0.759	0.146	1.066	–
Personality: conscientiousness	0.408	0.904	0.047	1.271	–
Personality: extraversion	0.409	0.912	-0.008	1.283	–
Personality: agreeableness	0.352	0.724	0.222	1.041	–
Personality: neuroticism	0.444	0.909	0.135	1.293	–
Semantic PC space	0.378	0.386	0.275	0.347	0.580
Style PC space	0.308	0.328	0.031	0.327	0.636

Table 7: Robust descriptive decomposition of RQ1 conversational trajectory geometry. Dashes denote scalar spaces in which orthogonal displacement is identically zero.

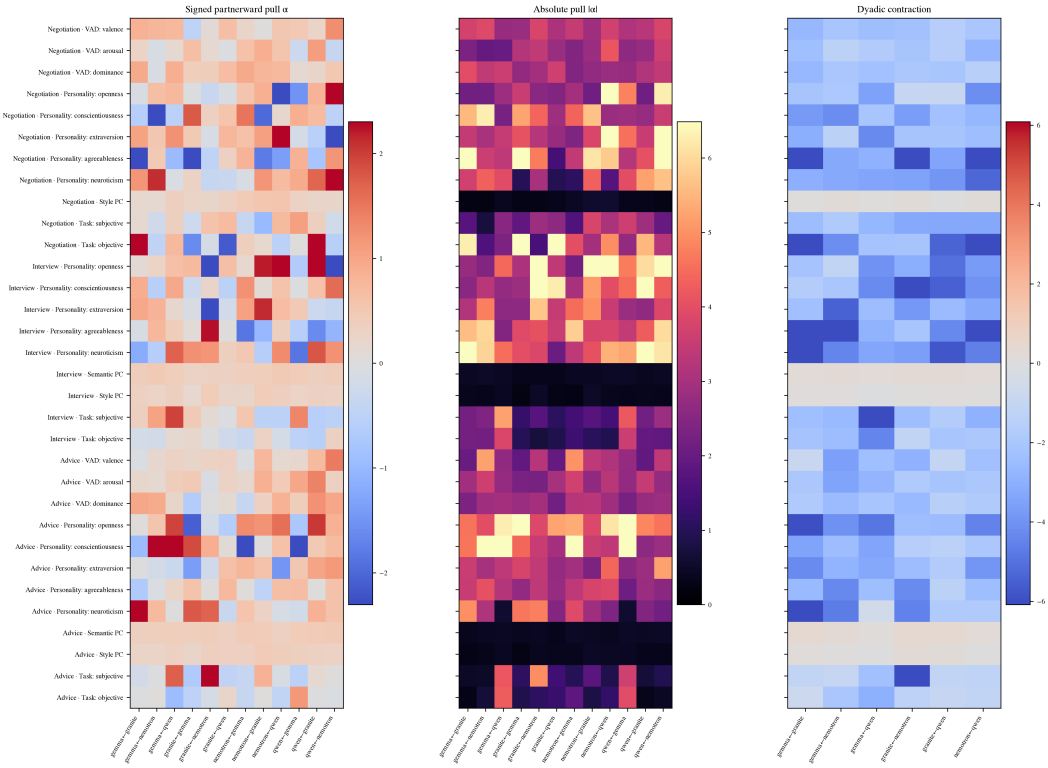


Figure 7: Dimension-level geometry in advice, interview, and negotiation by model dyad. Arrows denote speaker ← partner. All three panels exclude matched-reference distances below 10^{-4}; color limits use the 95th percentile of absolute cell values to prevent isolated ratios from compressing the remaining scale. Debate dimension-level geometry is reported in Appendix E.

Setting	Family	Self-play inertia	Mixed-play inertia
Negotiation	VAD	0.257	0.322
Negotiation	Personality	0.376	0.493
Negotiation	Style	0.201	0.264
Negotiation	Task	0.331	0.375
Interview	Personality	0.469	0.487
Interview	Semantic	0.121	0.174
Interview	Style	0.217	0.239
Interview	Task	0.616	0.511
Advice	VAD	0.457	0.415
Advice	Personality	0.460	0.491
Advice	Semantic	0.156	0.220
Advice	Style	0.223	0.258
Advice	Task	0.516	0.455

Table 11: Decoded-state inertia by conversation condition.

Setting	Family	$q = 0$	$q = .25$	$q = .50$	$q = .75$	$q = 1$
Negotiation	VAD	0.169	0.553	0.203	0.787	0.453
Negotiation	Personality	-0.120	0.220	0.433	-0.120	0.508
Negotiation	Style	0.253	0.296	0.322	0.345	0.347
Negotiation	Task	-0.510	0.041	0.604	0.286	0.421
Interview	Personality	0.229	-0.113	0.473	0.279	0.501
Interview	Semantic	0.279	0.369	0.401	0.417	0.433
Interview	Style	0.225	0.304	0.309	0.339	0.363
Interview	Task	0.666	-0.296	-0.194	0.316	0.030
Advice	VAD	0.182	0.850	0.350	0.277	0.446
Advice	Personality	0.152	0.085	0.462	0.831	0.460
Advice	Semantic	0.294	0.405	0.428	0.428	0.420
Advice	Style	0.243	0.305	0.337	0.370	0.375
Advice	Task	0.012	-0.116	-0.005	-0.261	0.880

Table 12: Denominator-filtered partnerward pull over relative conversation progress.

Table 13: Denominator-filtered partnerward pull by role direction.

Setting	Family	Speaker role	Partner role	Mean pull
Negotiation	VAD	role_a	role_b	0.427
Negotiation	VAD	role_b	role_a	0.439
Negotiation	Personality	role_a	role_b	0.178
Negotiation	Personality	role_b	role_a	0.190
Negotiation	Style	role_a	role_b	0.270
Negotiation	Style	role_b	role_a	0.356
Negotiation	Task	role_a	role_b	0.463
Negotiation	Task	role_b	role_a	-0.126
Interview	Personality	candidate	interviewer	0.129
Interview	Personality	interviewer	candidate	0.419
Interview	Semantic	candidate	interviewer	0.403
Interview	Semantic	interviewer	candidate	0.357
Interview	Style	candidate	interviewer	0.317
Interview	Style	interviewer	candidate	0.298
Interview	Task	candidate	interviewer	0.030
Interview	Task	interviewer	candidate	0.179
Advice	VAD	seeker	supporter	0.174
Advice	VAD	supporter	seeker	0.668
Advice	Personality	seeker	supporter	0.181
Advice	Personality	supporter	seeker	0.615
Advice	Semantic	seeker	supporter	0.380
Advice	Semantic	supporter	seeker	0.410
Advice	Style	seeker	supporter	0.331
Advice	Style	supporter	seeker	0.321
Advice	Task	seeker	supporter	0.092
Advice	Task	supporter	seeker	0.113

Table 14: Immediate efficacy and final targeted-variable ordering by intervention position. Relative efficacy is reported as median [IQR] across the 56 model–variable cells.

Position	Relative efficacy	Position winners	Ordered at target 10
T1	1.00 [0.92, 1.00]	36/56	42/56
T3	0.81 [0.57, 0.93]	8/56	45/56
T6	0.65 [0.31, 0.90]	5/56	41/56
T8	0.54 [0.23, 0.83]	3/56	41/56
T10	0.53 [0.21, 0.83]	4/56	—

G Additional Steering Details and Results

G.1 Complete position-dependent results

The main longitudinal evaluation considers interventions at target responses 1, 3, 6, and 8. We additionally evaluate target response 10 as an immediate-controllability diagnostic. Because no subsequent target response remains, this condition is included in the immediate positional analysis but excluded from endpoint-persistence and post-intervention trajectory comparisons. Immediate-efficacy normalization is computed across all five tested positions.

For each model–variable cell, let $\bar{y}_\lambda(p)$ denote the mean score of the steered target response across evaluation conversations when steering strength λ is applied at target position p. We define the immediate high–low effect as

$$\Delta_{\mathrm{imm}}(p) = \bar{y}_{+4}(p) - \bar{y}_{-4}(p). \tag{2}$$

Because score scales differ across model–variable cells, we normalize the effect at each position by the largest effect observed within that cell:

$$\widetilde{\Delta}_{\mathrm{imm}}(p) = \frac{\Delta_{\mathrm{imm}}(p)}{\max_{p'} \Delta_{\mathrm{imm}}(p')}. \tag{3}$$

The complete results are reported in Table 14.

G.2 Limited prompt-steering comparison

We compare activation steering with a fixed natural-language prompt baseline using Gemma-4-12B-IT self-play in four preselected context–variable settings (Table 15). Each setting uses ten evaluation starts, drawn from 5, 10, 10, and 7 source conversations for debate, advice, interview, and negotiation, respectively. Conditions share the first ten generated replies and matched generation seeds. Each continuation contains twenty new alternating replies in total. Only target response 6 (overall reply 11) is intervened upon, using activation strengths $\lambda = \pm 4$ or a high/low natural-language instruction appended to the original system prompt for that response only. The instruction is removed thereafter; generated replies remain in the conversation history. An unsteered condition is also retained.

For each method, we report the mean high-minus-low target-variable score at the intervention response (T6) and final target response (T10). Scoring uses label-derived MiniLM projections for stance, extraversion, and task adherence, and GoEmotions-derived valence. These are descriptive comparisons of one fixed prompt pair per variable against the selected activation strengths, without matching intervention intensity or tuning prompts.

Activation steering produces larger immediate gaps in all four settings. Endpoint effects are smaller and do not uniformly favor activation steering.

G.3 Partnerward alignment and sensitivity analyses

Because $d_{\min}$ is computed over the complete trajectory, its minimum may precede the intervention, particularly for later steering positions. We therefore interpret closest-distance advantage as a descriptive whole-trajectory diagnostic rather than exclusively as a post-intervention effect.

Table 15: Prompt baseline comparison. Entries are activation / prompt mean high–low score gaps. Positive values indicate the intended ordering. Scores retain variable-specific units and should only be compared within rows; they are not the position-normalized effects reported in Appendix G.1.

Context	Variable	Immediate (T6)	Final (T10)
Debate	Objective stance	$0.912/0.013$	$0.057/-0.126$
Advice	Valence	$0.531/0.089$	$0.068/-0.004$
Interview	Extraversion	$2.737/0.121$	$0.138/0.066$
Negotiation	Objective task adherence	$2.650/-0.149$	$0.247/0.508$

Table 16: **Position-dependent effects of one-shot steering.** Context-level medians across model–variable cells at target responses T1, T3, T6, and T8. Immediate efficacy is normalized within each cell; positive longitudinal advantages favor partnerward steering. Bold marks the largest value within each context and measure. T10 is omitted because no post-intervention target response remains.

Context	Measure	T1	T3	T6	T8
Debate	Immediate efficacy	**1.000**	0.630	0.230	0.223
	Closest-distance advantage	**+0.022**	+0.016	+0.002	+0.000
	Final contraction advantage	+0.231	**+0.419**	+0.060	+0.081
Advice	Immediate efficacy	**1.000**	0.761	0.466	0.229
	Closest-distance advantage	-0.007	+0.008	**+0.034**	+0.002
	Final contraction advantage	-0.025	**+0.061**	+0.055	+0.028
Interview	Immediate efficacy	**0.998**	0.853	0.855	0.779
	Closest-distance advantage	-0.003	**+0.011**	+0.001	+0.000
	Final contraction advantage	**+0.003**	-0.002	-0.006	-0.002
Negotiation	Immediate efficacy	**0.861**	0.846	0.827	0.718
	Closest-distance advantage	+0.002	**+0.026**	-0.007	+0.000
	Final contraction advantage	**+0.104**	**+0.104**	-0.044	-0.140

Table 16 reports the complete context-level positional summary corresponding to the condensed results in Table 1.

G.4 Complete generation-quality results

Table 17 reports the native-model likelihood and cross-turn repetition results for each generator model. Changes are measured relative to the matched unsteered condition. Positive $\Delta\mathrm{NLL}$ indicates lower likelihood under the original generator, while positive Δ repetition indicates increased cross-turn repetition.

Because the models use different native likelihood scales and tokenizations, raw NLL values should be compared only within a model; the within-model $\Delta\mathrm{NLL}$ values are the relevant steering contrasts.

G.5 Qualitative steering examples across models

Figure 8 and the following panels provide one illustrative example per generator model. Each example compares matched low-pole, unsteered, and high-pole responses generated from the same preceding conversation history. Examples were selected for large within-cell separation, bidirectional score ordering, and readability; they illustrate the intervention behavior rather than providing additional aggregate evidence.

Shared preceding partner response. *"That proactive approach to communication with your supervisor sounds like it would really help prevent small issues from snowballing into larger problems. When you're discussing those potential adjustments or extra support with a supervisor, how do you typically prefer to approach that conversation?"*

Low-pole steering	**Unsteered**	**High-pole steering**
$\lambda = -4$; score $= -0.989$	$\lambda = 0$; score $= 0.463$	$\lambda = +4$; score $= 1.724$
"I prefer to approach these conversations by first organizing a list of "known unknowns" and quantifying the exact impact on the deadline to ensure I was not presenting a problem without a proposed solution."	*"I prefer to approach those conversations with a lot of preparation, usually by bringing a clear summary of the current status and some potential solutions for the supervisor to consider."*	*"I would approach that conversation with a lot of sincerity, as I really value open and supportive communication. I'd make sure to share my perspective on how we can work together to navigate things, and I'd appreciate the opportunity to hear your thoughts too."*

Figure 8: **One-shot agreeableness steering produces a visible response shift in Gemma 4.** Excerpts from an illustrative conversation with steering at target response 6. The three conditions share the preceding history and satisfy $y_{+4} > y_0 > y_{-4}$. Scores are variable-specific and are not comparable across panels.

Shared preceding partner response. *"Absolutely, I am delighted to formalize this agreement. By exchanging 2 units of firewood for 3 units of food and adding 1 unit of water to support your hydration needs, we create a mutually beneficial outcome that aligns with our respective priorities."*

Low-pole steering	**Unsteered**	**High-pole steering**
$\lambda = -4$; score $= -1.151$	$\lambda = 0$; score $= 1.815$	$\lambda = +4$; score $= 2.490$
"With great enthusiasm, I accept this agreement wholeheartedly. As we partake in the splendor of nature, may our shared resources fuel both body and soul, weaving a tapestry of memories under the vast sky."	*"Certainly! With this agreement in place, I propose we proceed immediately to the handover of resources. My group will transfer 2 units of firewood and 1 unit of water to your camp, while you provide us with 3 units of food."*	*"Certainly! I'm pleased to confirm that we will proceed with the agreed exchange: my group will transfer 2 units of firewood and 1 unit of water to your group in return for 3 units of food."*

Figure 9: **One-shot task adherence (subjective) steering produces a visible response shift in Granite 4.1.** Excerpts from an illustrative conversation with steering at target response 3. The three conditions share the preceding history and satisfy $y_{+4} > y_0 > y_{-4}$. Scores are variable-specific and are not comparable across panels.

Shared preceding partner response. *"That's a great balance—adapting without losing sight of individual strengths. Could you share an example where reframing a role or task led to unexpected positive outcomes for the team or that person specifically?"*

Low-pole steering $\lambda = -4$; score $= -0.523$	**Unsteered** $\lambda = 0$; score $= 0.865$	**High-pole steering** $\lambda = +4$; score $= 1.551$
"In a past role, I noticed a team member thrived on strategic analysis but was often assigned repetitive administrative tasks. To align their strengths with the project's needs, I reframed their responsibilities to focus more on data-driven decision-making for campaign planning."	*"Absolutely! A few years ago, a teammate who loved creative storytelling but struggled with data-heavy tasks was assigned to help analyze campaign performance."*	*"Absolutely! A few years ago, we had a designer who was super creative but struggled with data-heavy tasks. Instead of forcing them into analytics, I shifted their role to lead the storytelling part of our campaigns—turning user insights into compelling narratives."*

Figure 10: **One-shot extraversion steering produces a visible response shift in Nemotron Nano.** Excerpts from an illustrative conversation with steering at target response 3. The three conditions share the preceding history and satisfy $y_{+4} > y_0 > y_{-4}$. Scores are variable-specific and are not comparable across panels.

Shared preceding partner response. *"That quiet strength settling into your bones is exactly where you need to be right now, and I am so glad we could find this gentle space together. You are already doing the hardest part by simply existing without the need to perform or generate constant energy."*

Low-pole steering $\lambda = -4$; score $= 0.296$	**Unsteered** $\lambda = 0$; score $= 0.729$	**High-pole steering** $\lambda = +4$; score $= 0.794$
"Hearing you say my support remains steady gives me a fragile hope that maybe I can actually rest without falling apart again."	*"Hearing you say my support remains steady gives me a sense of safety I haven't felt in so long, and it's okay to just let that quiet strength settle into my bones without rushing back to the noise."*	*"Hearing you say my support remains steady gives me a profound sense of safety that I haven't felt in a long time. It feels incredibly freeing to simply exist here without the pressure to constantly generate energy or fill every moment with noise."*

Figure 11: **One-shot advice valence steering produces a visible response shift in Qwen 3.5.** Excerpts from an illustrative conversation with steering at target response 6. The three conditions share the preceding history and satisfy $y_{+4} > y_0 > y_{-4}$. Scores are variable-specific and are not comparable across panels.

Table 17: Generation-quality results for one-shot activation steering. Changes are reported relative to matched unsteered trajectories.

Model	Base NLL	ΔNLL (-4)	ΔNLL $(+4)$	Base repetition	Δ repetition (-4)	Δ repetition $(+4)$
Qwen	1.3259	+0.0039	-0.0129	0.1453	-0.0021	+0.0081
Granite	0.5412	+0.0304	+0.0056	0.2872	-0.0316	-0.0080
Gemma	1.3008	+0.0933	+0.0788	0.1305	-0.0187	-0.0147
Nemotron	0.8242	+0.0027	+0.0277	0.4566	-0.0001	-0.0056