

Are LLM Agents Behaviorally Coherent? Latent Profiles for Social Simulation

James Mooney^{1*}, Josef Woldense^{2*}, Zheng Robert Jia¹, Shirley Anugrah Hayati¹,
My Ha Nguyen³, Vipul Raheja^{4†}, Dongyeop Kang¹

¹Department of Computer Science and Engineering, University of Minnesota

²Department of African American & African Studies, University of Minnesota

³Department of Sociology, University of Chicago, ⁴Google DeepMind

{moone174, jwoldens, dongyeop}@umn.edu

Abstract

The impressive capabilities of Large Language Models (LLMs) raise the possibility that synthetic agents can serve as substitutes for real participants in human-subject research. To evaluate this claim, prior research has largely focused on whether LLM-generated survey responses align with those produced by human respondents whom the LLMs are prompted to represent. In contrast, we address a more fundamental question: Do agents maintain *empirical consistency*; aligning to human behavioral models when examined under different experimental settings? To this end, we develop a study designed to (a) ask a set of questions which reveals an agent’s latent profile and (b) examine agent behavioral consistency in a conversational setting with other agents. This design enables us to explore a set of behavioral hypotheses to assess whether an agent’s conversational behavior is consistent with what we would expect from its revealed state. Our findings show significant inconsistencies in LLMs across model families and at differing model sizes. Most importantly, we find that, although agents may generate responses matching those of their human counterparts, they fail to be empirically consistent, representing a critical gap in their capabilities to accurately substitute for real participants in human-subject research.

1 Introduction

LLMs have demonstrated remarkable progress in recent years, prompting researchers and practitioners alike to ask not whether these systems can pass the Turing test (Jones & Bergen, 2025), but whether they can convincingly adopt full-fledged human personas (Hu & Collier, 2024; Park et al., 2023). Early findings suggest they can. For example, Park et al. (2024) find that when agents are constructed using rich qualitative interview data, they exhibit attitudes and behaviors that closely mirror those of their human counterparts. Such results have inspired what we term the *substitution thesis*: if agents can emulate humans, they may serve as substitutes for real participants in human-centered research. As substitutes, agents can be examined for individual traits or can be deployed to simulate human societies at scale. Should this prove viable, the potential upsides for social research would be tremendous: companies might test new products on virtual customers (Xiang et al., 2024; Ilagan et al., 2024), and social scientists could explore complex phenomena like war (Hua et al. (2024), governance Piatti et al. (2024), or cultural evolution Perez et al. (2024) with fewer ethical and logistical constraints.

Still, the gap between technological promise and practical utility remains large. While Park et al. (2024) achieve impressive persona fidelity, it does so by relying on lengthy two-hour interviews. In studies where agents are not given such extensive background information, their persona mimicry begins drifting in significant ways (Li et al., 2024; Abdulhai et al., 2025; Sandhan et al., 2026). For instance, when tasked with representing different American sociopolitical groups, LLM agents broadly matched aggregate human opinions but displayed far less variance, raising doubts about their use in downstream

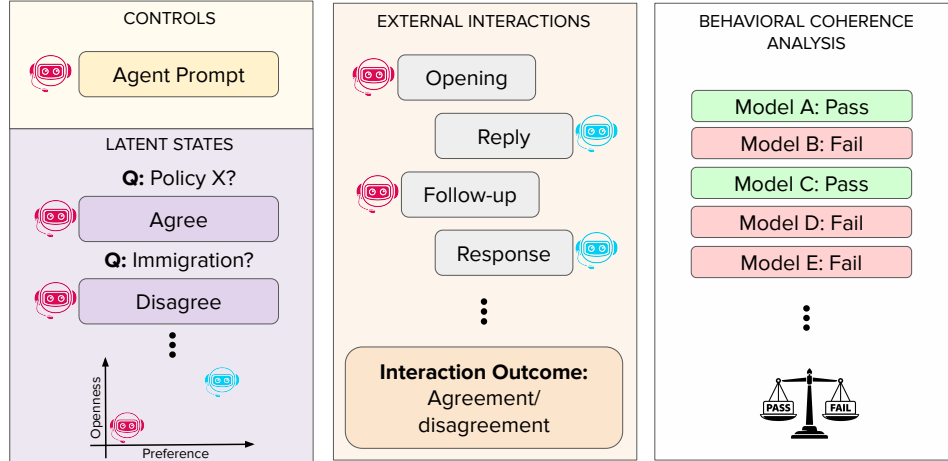


Figure 1: A high-level overview of our experimental framework. **Upper Left:** We prepare language model agents with variation and direct *Control* via prompting. **Bottom Left:** We ask agents individual questions with categorical responses to construct *Latent Profiles* (e.g., topic preferences, openness to new experiences). **Middle:** We pair agents and have them converse on various topics (*External Interactions*), measuring outcomes such as agreement over the course of a conversation. **Right:** We use the *Controlled* prompting inputs, *Latent States* (agree or disagree) from individual questions, and *External Interactions* from conversations to test against existing human behavioral models, expecting agents to behave consistently across all evaluation variables.

analyses (Bisbee et al., 2024). Similar “flattening effects” have also been observed across identity groups, where agent responses appear more homogeneous than their real-world counterparts (Wang et al., 2025a). We make further strides in identifying current gaps in the substitution thesis with the following contributions.

Proposed Framework. Our design rests on two pillars: (1) Collect a large amount of question-answer responses and conversational data for a controlled set of agents, and (2) testing whether the data collected in (1) matches the results expected from existing human behavioral models. If the collected data matches the results, then we define the language model as being *Consistent* over its inputs for the given behavioral model in question. Expanding on the first pillar of this framework, we separate the data collection into three distinct categories: (1) *Control Variables* which are introduced by systematically varying the system prompts of the agents considered, (2) *Latent States*, which are calculated by aggregating the responses of LLMs to simple, categorical questions relating to topic preference and general personality traits, and (3) *External Interaction* data, which are computed based on multi-turn conversations between different agents. Figure 1 demonstrates this more explicitly.

Findings. We explore this framework across six separate human behavioral models - looking to examine how smaller language models perform in a variety of settings. Our results reveal a persistent behavioral inconsistency. At an aggregate level: pairs which show larger preference gaps (i.e. one agent strongly agrees with a statement while another strongly disagrees) tend to reach higher agreement, and more open pairs tend to show higher agreement as well. However, these broad trends do not survive stricter tests of empirical coherence. Across model families, agents rarely sustain outright disagreement even when their stated preferences are maximally opposed; biasing prompts do not reliably restore principled disagreement; shared negative sentiment produces weaker alignment than shared positive sentiment; and openness loses predictive force in the very cases where it should matter most. In other words, current LLM agents can reproduce the appearance of coherent social behavior in the aggregate while failing more demanding checks that require their conversations to faithfully realize their own stated internal traits. Put plainly - as the behavioral models become more complex or more fine-grained, LLMs begin to fail at the task of replicating human behavior.

2 Related Work

LLMs as Agents and Human Substitutes in Dialogue Recent advances in Large Language Models (LLMs) have opened new possibilities for simulating human subjects in social science research. These models exhibit context-sensitive reasoning and structured decision-making capabilities (Wei et al., 2022; Kojima et al., 2022), enabling researchers to utilize them not only as tools but as experimental subjects (Mou et al., 2024; Park et al., 2023). In multi-agent simulations, LLMs have demonstrated socially emergent behaviors—forming memories, goals, and interaction patterns resembling real-world dynamics (Wang et al., 2025b). They have been used to model phenomena like conformity, information cocoons (Anthis et al., 2025), war (Hua et al., 2024), and market competition (Zhao et al., 2024). In structured survey settings, their responses have shown high alignment with human data across various conditions (Anthis et al., 2025). Nonetheless, significant conceptual and technical challenges remain. LLMs rely on statistical prediction rather than cognitive reasoning, and while they may appear behaviorally plausible, this can obscure underlying instability. They often fail to reproduce human-like distributional variance or demographic nuance and remain highly sensitive to prompt design and temporal drift (Bisbee et al., 2024; Petrov et al., 2024; Takata et al., 2024).

Behavioral Consistency among Personality, Preference, and Topic Although LLMs can maintain fluent conversation, they frequently lack continuity in personality and preference across multiple turns. Benchmarks like Topic-Conversation Relevance (TCR) assess topic relevance (Fan et al., 2024), but do not account for how personality traits might influence topic engagement or behavioral adaptation. Similarly, Long-Term Memory (LTM) benchmarks show that while LLMs can recall factual details, they struggle to retain identity- or preference-linked information over time (Castillo-Bolado et al., 2024).

Traditional persona-based models Zhang et al. (2018); Rashkin et al. (2019) allow for stylized variation (e.g., “likes cats”), but do not simulate evolving personality states or trait-informed reasoning. Recent works on generative agents with memory and reflection (Park et al., 2023) and trust-aware simulations (Xie et al., 2024) have made progress toward this goal, yet fall short in capturing how personality shapes topic alignment in dynamic conversations. Persona injection has been shown to improve coherence and emotional nuance (Wu et al., 2025). Trait-grounded personas help LLMs maintain consistent behaviors, influencing both the form and distribution of emotional support strategies. Synthetic datasets built from large-scale simulations further show that persona conditioning enhances diversity across psychological traits (Ge et al., 2024; Wu et al., 2025). However, challenges still remain, as studies have shown that dialogues generated without personas tend to be more concentrated and less diverse in psychological traits. In contrast, persona-conditioned outputs distribute more broadly across trait dimensions, such as Emotionality and Openness (Wu et al., 2025). The Big Five traits, including openness, are both stable across time and life events (Cobb-Clark & Schurer, 2012) and significantly correlated with resilience, cognitive flexibility, and adaptive functioning (Oshio et al., 2018).

3 A Framework for Probing Behavioral Coherence

To measure behavioral consistency, we design a novel experimental framework that probes how latent profiles of LLMs evolve through prompting and interaction. Rather than aiming to exhaustively characterize all possible behaviors, this framework serves as a prototype for systematically evaluating a specific class of behavioral dynamics. In particular, it captures how model-internal dispositions shift and manifest across conversational sequences.

3.1 Experimental Framework

At a high level, our framework’s tests undergo five sequential stages as shown in Figure 2: (1) select a topic, (2) generate agents, (3) elicit their Latent States, (4) pair them for dialogue, and (5) evaluate conversational agreement (details in Appendices A and B).

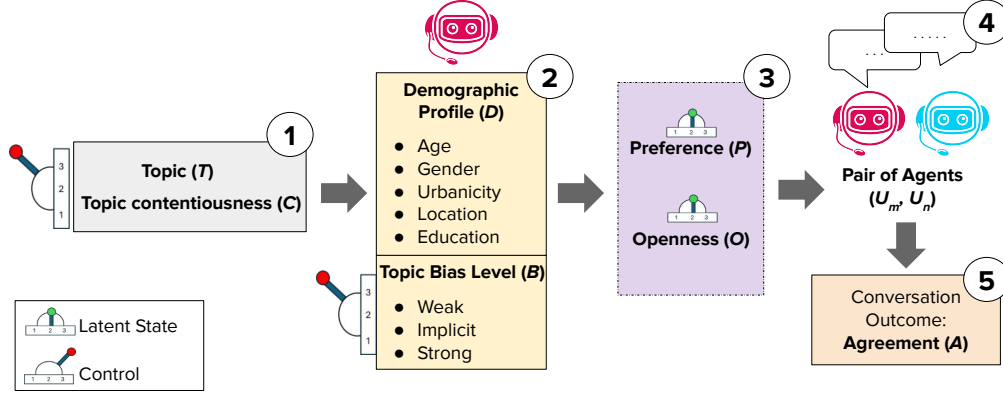


Figure 2: Proposed framework for probing behavioral coherence : (1) We first gather a set of topics of varying contentiousness levels to query agents on. (2) For a given topic, explicit agent profiles are gathered by varying the prompt among different demographic values (age, gender, etc.). This prompt is further altered to include information specific to the agent’s bias toward the chosen topic. (3) Latent states are gathered for each agent by asking a question about their preference on the given topic and about their openness to being swayed by others. (4) Agents are paired together to discuss the topic, and (5) agreement scores are calculated for each turn of their conversations.

Topic Selection: Construct a set of topics, where each topic T is associated with a contentiousness level $C \in \{1, 2, 3\}$, with 1 being the least contentious and 3 being the most contentious. The set contains nine topics in total, with exactly three topics assigned to each contentiousness level. Further descriptions of the topics may be found in Table 2.

Generate Agents: For each topic, construct agents with demographic profiles D_i defined by age, gender, urbanicity, location, and education (see Appendix B for more details on specific prompt construction). We further modify the agent prompt to include their bias towards the topic at hand, $B \in \{1, 2, 3\}$, with 1 being the least biased towards an opinion on a topic, and 3 being the most biased. Demographic region is limited to the United States and systematically varied across 5 age groups, 2 genders, 4 regions, 4 urbanicity levels, and 6 education levels.

Internal State: Identify each agent’s preference P_i and openness O_i . Both these values are captured by posing a set of questions to agent i and the responses are used to create a number indicating an agent’s preference for a topic and their openness - both of which are represented as integers between 1 and 5. See Section A for more details how these values are computed.

Pairing Agents: In order to properly test a given set of behavioral models, we must ensure that we are sampling from a large space of agent pairings during conversation. To this end, we must provide a process mechanism for pairing agents in the *External Interactions* portion of our framework.

As each agent i has a unique preference, openness and bias associated with it, the *controlled* profile of an agent may be given by (P_i, O_i, B_i) . Let $U := \{(P_j, O_j, B_j)\} \forall j \in \{1, \dots, N\}$ be the set of all agents. Then, for all possible pairs (U_m, U_n) , we sample agents q, r such that $(P_q, O_q, B_q) = U_m$ and $(P_r, O_r, B_r) = U_n$. This means that we are mechanically preparing all possible combinations of preference, openness and bias to interact with one another - enabling higher fidelity when mapping to our behavioral models of interest.

Conversation Outcome: For each step K of a conversation, we use LLM-as-judge to score agreement $A \in \{1, 2, 3, 4, 5\}$ (1 = complete disagreement, 5 = complete agreement). For analysis, we retain only the final agreement score. To calibrate judgments, we provide five annotated sample conversations—one for each score.

3.2 Specific Experiments

We examine 9 topics with different controversy levels to explore the effect of contentiousness on both individual agent response as well as on agreement during interaction with other agents. Among these topics (in descending order of contentiousness) are opinions on "Taxes", "Remote Work" and "Coca Cola vs. Pepsi". More details on specific topics and agent demographics can be found in Appendix A (in particular, Tables 2 and 3 show the set of topics and agent demographics we consider).

Using these topics and demographics, we conduct two sets of experiments: (1) qualitative experiments (§4.1 and §4.2) as a proof-of-concept on LLMs’ internal consistency on preferences and openness, and (2) robustness experiments (§4.3) to provide statistical comparisons across various LLM model sizes. For both experiment sets, we use Qwen3-32B (Yang et al., 2025) as the judge. For all the qualitative visualizations below, we use Gemma-3-12b-it (Team et al., 2025) as the agent model. For the purposes of robustness, we examine the results of our approach across different sizes of models for the Gemma3 (Team et al., 2025), Llama3.x (et al., 2024), and Qwen3 (Yang et al., 2025) model families.

4 Experimental Results

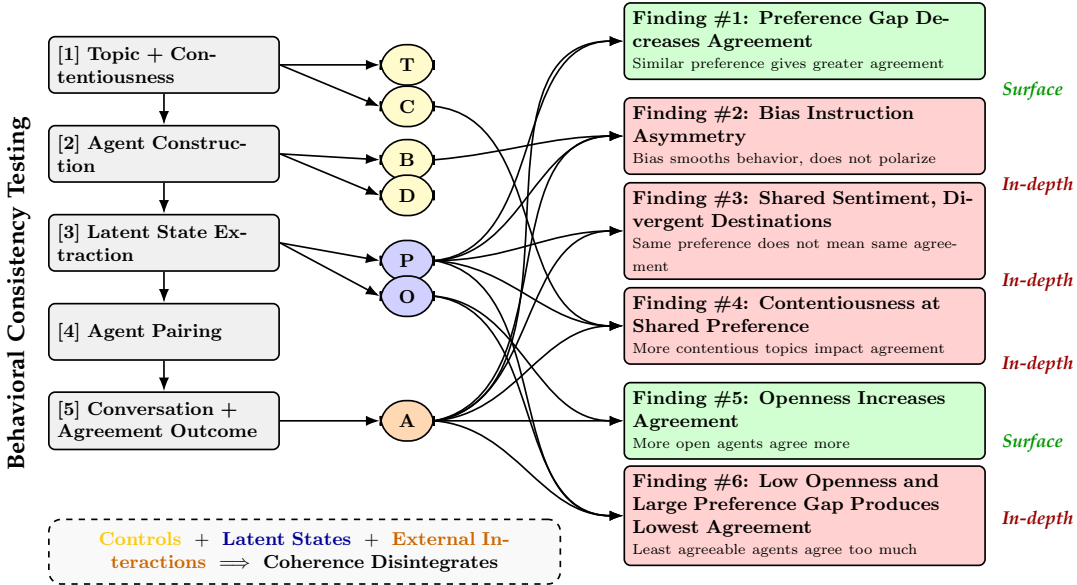


Figure 3: A summary of main findings aligned with our pipeline. Our pipeline is presented at left and shows the corresponding artifacts produced in the center panel (e.g. T, C, B , etc.). One may refer to Figure 2 for a more in-depth visualization of this process. With these **control variables**, **latent states** and **External Interaction** artifacts collected, we test our approach on a variety of human behavioral models to see if LLMs can accurately match human behavior and retain their consistency. We focus on 6 separate behavioral models (presented in greater detail in Section 4). We find that, when using more complex, granular behavioral models, LLMs begin to break down and are no longer coherent among the different **latent states** and **external evaluations** collected. In other words, *Surface*-level tests **pass**, whereas *In-Depth*-tests **fail**.

In our experiments, we aim to uncover the discrepancy between the *appearance* of behavioral consistency and its breakdown under closer examination with more complex tests. We organize our findings around **preference** and **openness** as these are common areas of study for behavioral models (Lupton & Thornton, 2017; Yokomatsu & Kotani, 2021; Lodge et al., 2006) and will enable straightforward evaluations when placed in our framework. We further break our results down by whether the test is "Surface-level" or "In-depth". This is an implicit dimension, but typically relies upon looking at our data in a more granular fashion

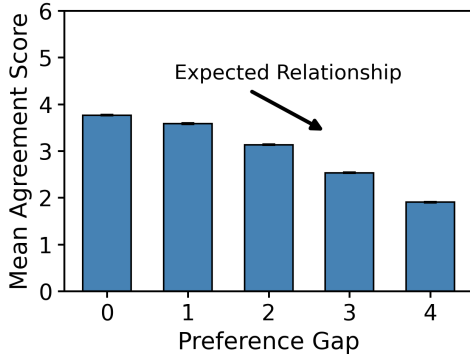


Figure 4: When two agents have a larger gap in their preference for a topic (e.g. favorite soda, views on taxes, etc.), they have less agreement in conversation.

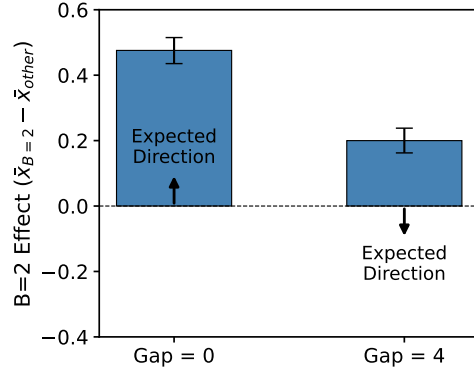


Figure 5: High bias leads to greater agreement in agents with a preference gap of 0, but does not lead to greater disagreement with a preference gap of 4.

- moving from simpler statistical tests to those which are more complex. To reiterate, we are using the Gemma-3-12b-it models for all visualizations that follow. Figure 3 generally demonstrates our results across the six behavioral models discussed in detail below. For more details on the precise nature of statistical tests discussed in this section, refer to Appendix C. We also provide a series of qualitative examples showing the agent profiles and conversation outputs for each of the findings in Appendix D.

4.1 On Preferences

Finding #1: Preference Gap Decreases Agreement We start with the the most basic consistency test: Agents with shared preferences should be more likely to agree than those with opposing preferences (Huckfeldt & Sprague, 1995; Mutz, 2006; Lupton & Thornton, 2017; Yokomatsu & Kotani, 2021; Stanley et al., 2020; Lodge et al., 2006). We measure this using the *preference gap* that is computed for each conversation pair. The latter is defined as the absolute difference between the agents’ preference scores. A gap of four indicates maximal divergence whereas a gap of zero indicates identical views.

Figure 4 shows that the results are consistent with our expectation: pairs with aligned preferences (gap=0) achieve the highest agreement, while pairs with greater gaps yield progressively lower scores. A Pearson correlation between preference gap and agreement score confirms a significant negative relationship ($r < 0, p < 0.01$), indicating that agreement decreases as the preference gap widens.

Finding #2: Bias Instruction Asymmetry In our framework, recall that the bias term is used to influence the relative strength of a belief around a given topic. In our case, the strongest bias is indicated by the bias parameter $B_i = 2$, which directs agents to adopt a strong stance either for or against the topic. We would expect the agents to reinforce both agreement *and* disagreement – amplifying agreement when agents share preferences and amplifying disagreement when the preferences diverge (Sherif & Hovland, 1961; Krosnick & Petty, 1995; Lord et al., 1979; Taber & Lodge, 2006; Kurmyshev et al., 2011).

We evaluate this by comparing conversations where both agents receive the bias instruction ($B_1 = B_2 = 2$) against all other configurations, which serve as the baseline, at two extremes of the preference gap. As illustrated in Figure 5, effects are interpreted relative to this baseline where positive values indicate higher agreement than the baseline, while negative values indicate lower agreement. The expected pattern is directional – at gap= 0, the bar should shift upward (above baseline), whereas at the maximum gap of four, it should shift downward (negative below baseline). At gap= 0, where agents share identical preferences, this expectation is borne out: bias amplifies agreement, with a one-sided Mann–Whitney U test confirming that $B = 2$ pairs achieve significantly higher agreement than the baseline ($p < 0.01$, Bonferroni-corrected for two comparisons).

At the maximum gap of four, where agents hold maximally divergent preferences, the expectation is the opposite. The bar should move in the negative direction to indicate reduced agreement relative to the baseline. However, as shown in Figure 5, this pattern does not materialize. Rather than scoring below the baseline, the agreement levels are actually higher – thus revealing a clear asymmetry in how bias instructions affect agent behavior across preference gaps.

Finding #3: Shared Sentiment, Divergent Destinations Consider two pairs of agents: one pair strongly favors a topic (preferences 5,5), while the other strongly dislikes it (preferences 1,1). In principle, both pairs should exhibit high agreement – one through shared enthusiasm, the other through shared aversion. Sentiment alignment, whether positive or negative, should yield comparable agreement (Abelson, 1968; Krosnick & Petty, 1995; Duck & Barnes, 1992; Kimura, 2003; Liu et al., 2023).

Contrary to expectation, Figure 6 reveals a striking directional bias in agreement outcomes. For nearly all gaps, high-aligned pairs ($\cdot, 5$) outperform the corresponding low-aligned pairs ($1, \cdot$), showing that agreement depends on the direction of alignment. Strikingly, despite having a gap of 0 and theoretically expected to achieve the highest agreement scores, pair (1,1) consistently scores below other pairs with larger gaps. These patterns are formally confirmed by one-sided Mann-Whitney U tests with Bonferroni correction, which show that low-aligned pairs rarely achieve significantly higher agreement than their high-aligned counterparts across the gap spectrum, highlighting a pronounced inconsistency in how shared sentiment translates into agreement.

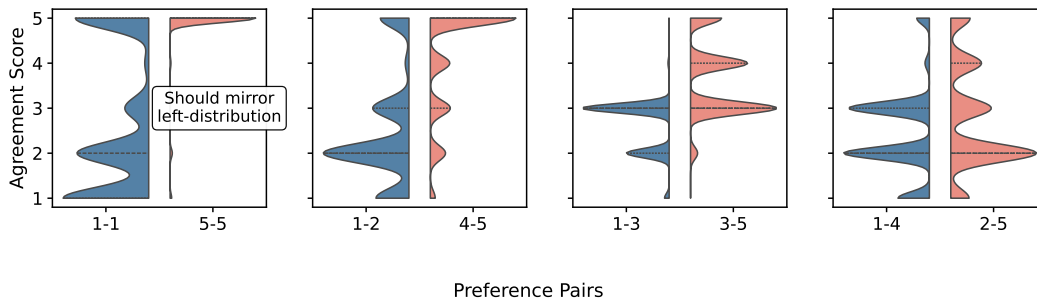


Figure 6: The distribution of agreement scores for fixed preference-gaps between two agents do not match. As an example, preferences of 1-1 and 5-5 in agents (left-most plot above) do not result in the same distributions of agreement scores, contrary to what behavioral models suggest it should.

Finding #4: Contentiousness at Shared Preference In principle, when both agents hold identical views on a topic, the inherent contentiousness of that topic should not independently affect agreement. Prior work in social psychology shows that interpersonal agreement is primarily driven by attitude similarity rather than the specific topic under discussion. Individuals who share the same evaluation of an issue tend to exhibit similar levels of agreement across topics with widely varying levels of social salience or controversy (Byrne, 1971; Krosnick & Petty, 1995; Huckfeldt & Sprague, 1995; Mutz, 2006; Klein et al., 2001).

Translated into the current study, we fix the preference gap at zero (i.e., both agents share the same preference) to isolate the role of topic characteristics. Our objective is to determine whether changes in topic contentiousness ($C = 1, 2, 3$) lead to systematic differences in agreement. Under the prior literature, the null hypothesis makes a clear prediction: when preferences are aligned, the distribution of agreement should remain effectively unchanged across levels of contentiousness – any observed differences should be negligible and attributable to random variation rather than the topic itself.

Contrary to this expectation, Figure 7 shows that, even with the preference gap held constant, agreement scores vary meaningfully across levels of contentiousness. This is formally

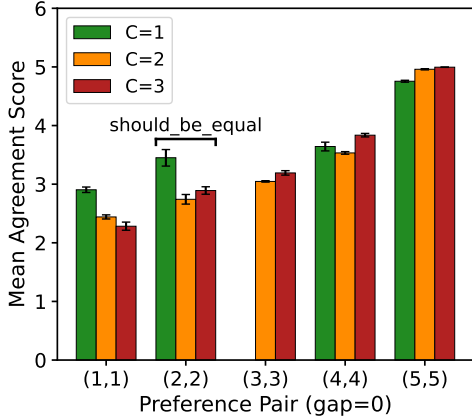


Figure 7: Agreement varies with topic controversy even for identical preferences, especially under strong disagreement.

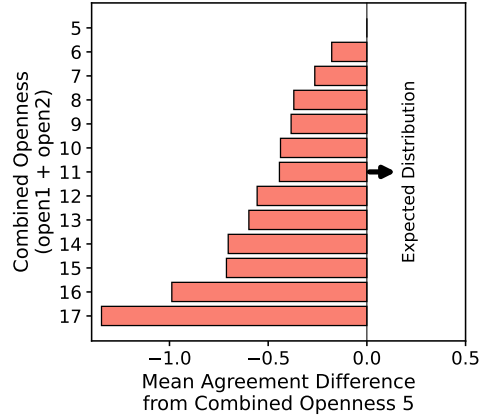


Figure 8: With maximal preference gaps, lower openness yields higher agreement.

corroborated using the Kruskal–Wallis test, which assesses whether agreement distributions differ across groups.

4.2 On Openness

Finding #5: Openness Increases Agreement At the aggregate level, the relationship between openness and agreement aligns with expectations. As the combined openness of a conversation pair increases, average agreement also rises, consistent with the intuition that more receptive agents converge more readily with their partners (Levitan & Visser, 2008; Petty & Cacioppo, 1986; Briñol & Petty, 2023; Robinson et al., 2024). We formalize this with a Pearson correlation between combined openness (the sum of both agents’ openness scores) and agreement. The test confirms a significant positive relationship ($r > 0$, $p < 0.01$), indicating that openness reliably increases agreement in aggregate. However, as with preferences, this surface-level finding masks deeper inconsistencies.

Finding #6: Low Openness and Large Preference Gap Produces Lowest Agreement If openness functions as intended, then the combination of low openness and large preference divergence should produce the lowest possible agreement – stubborn agents who disagree should be the hardest pair to reconcile (Lord et al., 1979; Wood & Quinn, 2003; Kunda, 1990; Wang et al., 2022; Hu et al., 2024). We test this by isolating conversations with maximal preference divergence (pairs 1,5) and comparing the agreement distribution of the lowest-openness pairing against all other openness pairings within this subset.

Specifically, we identify the openness pairing with the smallest combined openness score and test whether its agreement distribution falls significantly below each alternative pairing. We apply one-sided Mann–Whitney U tests with Bonferroni correction across all comparisons, requiring that the low-openness baseline produces significantly lower agreement than every other pairing ($p < 0.01$, corrected). Figure 8 shows that the lowest openness pairing results in models with the highest agreement scores.

Model	Preference				Openness	
	Surface		In-depth		Surface	In-depth
	T1	T2	T3	T4	T5	T6
Qwen3-0.6B	×	×	×	×	×	×
Qwen3-4B	×	×	×	×	✓	×
Qwen3-8B	✓	×	×	×	✓	✓
Llama-3.2-1B	×	×	×	×	✓	×
Llama-3.2-3B	✓	×	×	×	✓	×
Llama-3.1-8B	✓	✓	×	×	✓	×
gemma-3-1b	×	×	×	×	×	×
gemma-3-4b	✓	×	×	×	✓	×
gemma-3-12b	✓	×	×	×	✓	×

Table 1: Significance testing results across six human behavioral model tests applied to LLMs. ✓ denotes pass, × denotes failure.

4.3 Robustness across models

The qualitative findings above suggest systematic inconsistencies in the observed behaviors of LLMs. We apply these tests to other models to determine their robustness. A model is considered to *pass* a test if its outcomes align with the corresponding behavioral model. Appendix C defines the explicit statistical tests we use in more detail.

Table 1 shows the results of all tests across the set of models we considered. Overall, these results show that while models capture broad, surface-level trends (Tests 1 and 5), they systematically fail tests requiring deeper internal coherence. Importantly, this pattern holds across model sizes and families, suggesting that such inconsistencies are not idiosyncratic but general properties of current LLMs.

The following example shows one full behavioral trajectory of agents on contentiousness at shared preference (See Appendix D for more):

[1] Topic → [3] Internal State → [6] Outcome 8 FAIL

Finding 4 — Contentiousness Leakage: [1] Topic matters even when [3] Internal State is held identical across pairs. A more direct failure of Test 4 can be shown using exactly the preference pair $(P_1, P_2)=(1,1)$ across two conversations whose topics differ in contentiousness. Once the pair is fixed at shared negative sentiment, topic contentiousness should *not* independently drive the [6] Outcome .

— Low-contentiousness ($C=1$): spring vs. fall —

A1. Woman; twenties; suburban; Southern US; college; neg. spring bias. $P_1=1$ $O_1=5$.
A2. Woman; fifties; urban; Eastern US; some college; neg. spring bias. $P_2=1$ $O_2=6$.
[6] Outcome : $A = (2, 3, 5, 5)$.
A1: “No, I don’t. I like Fall a lot better.”
A2: “No, I don’t enjoy Spring. It’s too wet and muddy for my taste.”
A2: “Spring’s a mess, all them flowers bloomin’ and pollen everywhere. I like Fall, it’s cooler, leaves changin’ colors.”

— High-contentiousness ($C=3$): immigration —

A1. Woman; twenties; rural; Western US; some college; impl. neg. immigration bias. $P_1=1$ $O_1=6$. A2. Woman; sixties; exurban; Midwestern US; high school; expl. neg. immigration bias. $P_2=1$ $O_2=6$.
[6] Outcome : $A = (2, 2, 2, 2)$.
A1: “I think we’re takin’ in too many folks at once. It’s hard for our communities to keep up, you know?”
A2: “I hear ya. The border situation has gotten out of hand, and it’s causin’ real problems for regular Americans.”
A1: “Right, and there’s real safety concerns too.”
A2: “Exactly. We need tougher enforcement, or things will get worse.”

Both conversations share identical [3] Internal State values, yet the [1] Topic’s social charge produces divergent [6] Outcome trajectories — demonstrating that external context leaks directly through the pipeline.

5 Conclusion and Discussion

This paper set out to examine the substitution thesis: the idea that LLM agents might serve as substitutes for humans in social and behavioral research. Our contribution has been to shift the focus from singular or more monolithic benchmarks (i.e., alignment with human survey responses or demographic priors) to a more fundamental criterion: *empirical behavioral consistency*. Specifically, we asked whether LLM agents behave in ways that are coherent with respect to more complex, multi-agent behavioral baselines.

Our results reveal clear limitations in current LLM agents. While agents often appear consistent on the surface, closer inspection shows systematic deviations. Across settings, agents show a persistent inability to handle more complex or fine-grained modeling. These patterns were observed across model families and sizes, indicating they are not artifacts of

a single architecture but reflect broader limitations of current LLMs (at least for the scales we considered here). The implications are significant for using LLMs in social simulation and behavioral modeling. Although these systems can produce human-like responses in isolated cases, they fail to sustain trait-driven coherence across contexts, raising doubts about their reliability as stand-ins for real human participants.

Finally, while our current instantiation focuses on a single structured interaction setting, the framework is inherently extensible and can be adapted to a broader range of behavioral sequences and ordering schemes which remain as future work. Through a focused case study, we use this setup to examine the extent to which such behaviors remain stable or exhibit unreliability under controlled variations.

References

- Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. Consistently simulating human personas with multi-turn reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
- Robert P. Abelson. Theories of cognitive consistency: A sourcebook. In Gardner Lindzey and Elliot Aronson (eds.), *The Handbook of Social Psychology*, volume 2, pp. 112–139. Addison-Wesley, Reading, MA, 2 edition, 1968.
- Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. Llm social simulations are a promising research method. *arXiv preprint arXiv:2504.02234*, 2025.
- James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416, 2024.
- Pablo Briñol and Richard E. Petty. Openness and persuasion: Multiple processes, meanings, and outcomes. *Current Opinion in Psychology*, 49:101507, 2023. doi: 10.1016/j.copsyc.2022.101507.
- Donn Byrne. *The Attraction Paradigm*. Academic Press, New York, 1971. ISBN 9780121488506.
- David Castillo-Bolado, Joseph Davidson, Finlay Gray, and Marek Rosa. Beyond prompts: Dynamic conversational benchmarking of large language models. *arXiv preprint arXiv:2409.20222*, 2024.
- Deborah A Cobb-Clark and Stefanie Schurer. The stability of big-five personality traits. *Economics Letters*, 115(1):11–15, 2012.
- Steve Duck and Melanie K Barnes. Disagreeing about agreement: Reconciling differences about similarity. *Communications Monographs*, 59(2):199–208, 1992.
- Aaron Grattafiori et al. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Yaran Fan, Jamie Pool, Senja Filipi, and Ross Cutler. Topic-conversation relevance (tcr) dataset and benchmarks. *Advances in Neural Information Processing Systems*, 37:140159–140174, 2024.
- Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024.
- Hong-Xiang Hu, Guanghui Wen, Yun Chen, and Wei Xing Zheng. Opinion dynamics analysis for stubborn individuals in cooperation–competition networks based on path-dependence framework. *Automatica*, 169:111840, 2024.

- Tiancheng Hu and Nigel Collier. Quantifying the persona effect in LLM simulations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10289–10307, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.554. URL <https://aclanthology.org/2024.acl-long.554/>.
- Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. War and peace (waragent): Large language model-based multi-agent simulation of world wars, 2024. URL <https://arxiv.org/abs/2311.17227>.
- R. Robert Huckfeldt and John Sprague. *Citizens, Politics, and Social Communication: Information and Influence in an Election Campaign*. Cambridge Studies in Public Opinion and Political Psychology. Cambridge University Press, Cambridge, UK, 1995. ISBN 9780521030441. doi: 10.1017/CBO9780511664113.
- Joseph Benjamin Ilagan, Zachary Matthew Alabastro, Claire Basallo, and Jose Ramon Ilagan. Exploratory customer discovery through simulation using chatgpt and prompt engineering. 02 2024.
- Cameron R Jones and Benjamin K Bergen. Large language models pass the turing test. *arXiv preprint arXiv:2503.23674*, 2025.
- Yasuhiko Genku Kimura. Alignment beyond agreement. *The Journal of Integral Thinking*, 2003.
- Katherine J Klein, Amy Buhl Conn, D Brent Smith, and Joann Speer Sorra. Is everyone in agreement? an exploration of within-group agreement in employee perceptions of the work environment. *Journal of Applied Psychology*, 86(1):3, 2001.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- Jon A. Krosnick and Richard E. Petty. Attitude strength: An overview. In Richard E. Petty and Jon A. Krosnick (eds.), *Attitude Strength: Antecedents and Consequences*, pp. 1–24. Lawrence Erlbaum Associates, Mahwah, NJ, 1995. ISBN 9780805816182.
- Ziva Kunda. The case for motivated reasoning. *Psychological Bulletin*, 108(3):480–498, 1990. doi: 10.1037/0033-2909.108.3.480.
- Evguenii Kurmyshev, Héctor A Juárez, and Ricardo A González-Silva. Dynamics of bounded confidence opinion in heterogeneous social networks: Concord against partial antagonism. *Physica A: Statistical Mechanics and its Applications*, 390(16):2945–2955, 2011.
- Lindsey C. Levitan and Penny S. Visser. Reconsidering the strength of attitude strength: How attitude importance and attitude certainty influence attitude resistance to change. *Personality and Social Psychology Bulletin*, 34(7):939–952, 2008. doi: 10.1177/0146167208316699.
- Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Measuring and controlling instruction (in)stability in language model dialogs. In *COLM 2024: Conference on Language Modeling*, 2024. URL <https://arxiv.org/abs/2402.10962>.
- Xueyi Liu, Rui Hou, Yanglei Gan, Da Luo, Changlin Li, Xiaojun Shi, and Qiao Liu. Aspect-oriented opinion alignment network for aspect-based sentiment classification. *arXiv preprint arXiv:2308.11447*, 2023.
- Milton Lodge, Charles Taber, and Christopher Weber. First steps toward a dual-process accessibility model of political beliefs, attitudes, and behavior. In *Feeling politics: Emotion in political information processing*, pp. 11–30. Springer, 2006.
- Charles G. Lord, Lee Ross, and Mark R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098–2109, 1979. doi: 10.1037/0022-3514.37.11.2098.

- Robert N. Lupton and Judd R. Thornton. Disagreement, diversity, and participation: Examining the properties of several measures of political discussion network characteristics. *Political Behavior*, 39(3):585–608, 2017. doi: 10.1007/s11109-016-9371-7.
- Dan P McAdams. The five-factor model in personality: A critical appraisal. *Journal of personality*, 60(2):329–361, 1992.
- Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, et al. From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv preprint arXiv:2412.03563*, 2024.
- Diana C. Mutz. *Hearing the Other Side: Deliberative versus Participatory Democracy*. Cambridge University Press, Cambridge, UK, 2006. doi: 10.1017/CBO9780511617201.
- Shaul Oreg and Noga Sverdlik. Source personality and persuasiveness: Big five predispositions to being persuasive and the role of message involvement. *Journal of personality*, 82(3):250–264, 2014.
- Atsushi Oshio, Kanako Taku, Mari Hirano, and Gul Saeed. Resilience and big five personality traits: A meta-analysis. *Personality and individual differences*, 127:54–60, 2018.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST ’23*, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL <https://doi.org/10.1145/3586183.3606763>.
- Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative agent simulations of 1,000 people, 2024. URL <https://arxiv.org/abs/2411.10109>.
- Jérémy Perez, Corentin Léger, Marcela Ovando-Tellez, Chris Foulon, Joan Dussauld, Pierre-Yves Oudeyer, and Clément Moulin-Frier. Cultural evolution in populations of large language models. *arXiv preprint arXiv:2403.08882*, 2024.
- Nikolay B Petrov, Gregory Serapio-García, and Jason Rentfrow. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis. *arXiv preprint arXiv:2405.07248*, 2024.
- Richard E. Petty and John T. Cacioppo. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. Springer-Verlag, New York, 1986. ISBN 9780387963441. doi: 10.1007/978-1-4612-4964-1.
- Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems*, 37:111715–111759, 2024.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381, 2019.
- Michael D Robinson, Jade A Lindquist, and Roberta L Irvin. Open-hearted: Preferences for openness in the agreeableness domain. *Journal of Personality*, 92(2):378–392, 2024.
- Jivnesh Sandhan, Fei Cheng, Tushar Sandhan, and Yugo Murawaki. Persona jailbreaking in large language models. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 1412–1430, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.72. URL <https://aclanthology.org/2026.findings-eacl.72/>.

Muzafer Sherif and Carl I. Hovland. *Social Judgment: Assimilation and Contrast Effects in Communication and Attitude Change*. Yale University Press, New Haven, CT, 1961. ISBN 9780300002844.

Matthew L Stanley, Paul Henne, Brenda W Yang, and Felipe De Brigard. Resistance to position change, motivated reasoning, and polarization. *Political Behavior*, 42(3):891–913, 2020.

Charles S Taber and Milton Lodge. Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, 50(3):755–769, 2006.

Ryosuke Takata, Atsushi Masumori, and Takashi Ikegami. Spontaneous emergence of agent individuality through social interactions in llm-based communities. *arXiv preprint arXiv:2411.03252*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Keanealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Pettrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Pöder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Ütku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yu-vein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL <https://arxiv.org/abs/2503.19786>.

A. Wang, J. Morgenstern, and J.P. Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nat Mach Intell*, 2025a.

- Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, et al. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):1–37, 2025b.
- Lingfei Wang, Carmela Bernardo, Yiguang Hong, Francesco Vasca, Guodong Shi, and Claudio Altafini. Consensus in concatenated opinion dynamics with stubborn agents. *IEEE Transactions on Automatic Control*, 68(7):4008–4023, 2022.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Wendy Wood and Jeffrey M. Quinn. Forewarned and forearmed? two meta-analytic syntheses of forewarnings of influence appeals. *Psychological Bulletin*, 129(1):119–138, 2003. doi: 10.1037/0033-2909.129.1.119.
- Shenghan Wu, Yang Deng, Yimo Zhu, Wynne Hsu, and Mong Li Lee. From personas to talks: Revisiting the impact of personas on llm-synthesized emotional support conversations. *arXiv preprint arXiv:2502.11451*, 2025.
- Wei Xiang, Hanfei Zhu, Suqi Lou, Xinli Chen, Zhenghua Pan, Yuping Jin, Shi Chen, and Lingyun Sun. Simuser: Generating usability feedback by simulating various users interacting with mobile applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642481. URL <https://doi.org/10.1145/3613904.3642481>.
- Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, et al. Can large language model agents simulate human trust behavior? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.
- Muneta Yokomatsu and Hitomu Kotani. Knowledge sharing, heterophily, and social network dynamics. *The Journal of Mathematical Sociology*, 45(2):111–133, 2021.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, 2018.
- Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. Competeai: understanding the competition dynamics of large language model-based agents. In *Proceedings of the 41st International Conference on Machine Learning, ICMML'24*. JMLR.org, 2024.

A Generic Design Principles

The experimental design elicits internal states from agents and tests whether these states manifest consistently in dialogue. Below, we describe each component in turn.

Topic Contentiousness (C) It is commonly accepted that some topics are inherently more polarizing than others. A discussion on taxes is likely to provoke more disagreement than a conversation on the weather. And yet, any topic has the potential to polarize once the right mix of people is involved. From obscure debates over historical events to the football fans rooting for rival teams, seemingly innocuous subjects can spark intense disagreement when divergent viewpoints collide. By varying the contentiousness level of the topic at hand, we assess whether our agents exhibit consistent behavior—disagreeing when their preferences diverge, regardless of the subject matter. We assign each topic a contentiousness score $C \in \{1, 2, 3\}$, where 1 is least contentious and 3 is most.

Bias in Prompting (B_i) LLMs are known to exhibit sycophancy, often being overly agreeable to interlocutors. To counteract this, we introduce a bias parameter B_i that explicitly directs agents to take a stance.

- $B = 1$: No bias information added.
- $B = 2$: Implicit biasing (e.g., “You are a liberal Democrat” when the topic is immigration).
- $B = 3$: Explicit biasing (e.g., “You support immigration” when the topic is immigration).

For $B = 2$ and $B = 3$, agents are further directed to adopt either a positive or negative position on the topic.

Preference (P_i) Our primary measure of internal state is P_i , an agent’s preference on a given topic. The expectation is straightforward: preferences should predict conversational outcomes. Agents with aligned preferences should agree, while those further apart should be more likely to disagree. To elicit P_i , we prompt agents to take a position on statements such as “*taxes help to meet the needs of society*” or “*Coca-Cola is better than Pepsi*”. Responses are given on a 1–5 scale, where 1 indicates strong disagreement and 5 indicates strong agreement.

Openness (O_i) Given that the experiment takes place in a conversational setting, the outcome (i.e., level of agreement) will depend not only on the agents’ preferences, but also on their susceptibility to be swayed by their dialogue partner. To account for this, we draw on the concept of *Openness* from the Big Five personality framework, which is a trait linked to receptivity to new ideas and persuasiveness (McAdams, 1992; Oreg & Sverdlik, 2014). To make it suitable for our purpose, we modify the questions to capture the likelihood that an agent will revise its position when confronted with an opposing view. Denoted as, O_i , we measure openness by asking nine Yes/No questions, such as “*Do you often second-guess your choices after hearing someone else’s opinion?*” and “*Are you comfortable disagreeing with someone, even if they are a close friend or authority figure?*”. The additive index of responses produces an openness score: higher values indicate receptivity, while values near zero indicate rigidity.

Pairing Agents Once P_i and O_i are established, we assign agents into pairs for dialogue. An agent i is represented by its profile (P_i, O_i, B_i) . Pairings are constructed to maximize diversity, including aligned vs. opposed preferences and varying levels of openness. Measuring consistency at the level of pairs, rather than individuals, mitigates noise from idiosyncratic deviations. Group-level patterns thus provide a clearer signal of whether internal states predict conversational outcomes.

B Specific Framework

In this section, we detail the exact methodology (prompts, etc.) used within Section 3 of our paper.

(a) Topic Preparation Table 2 shows the set of topics we explore within our study. We examine several topics at each level of contentiousness in order to examine the effect

of contentiousness on both individual agent response as well as on agreement during interaction with other agents.

(b) Agent Construction with External Profiles As described in Section 3 of our original paper, agents are composed of both a demographic background (D) as well as information relating to their bias towards the topic of discussion (B).

The possible demographics of the agent are described in Table 3. An agent is composed of only one value from each trait category, and these values are used to construct the agent’s system prompt. For example, one such agent in our study would have as part of it’s system prompt *You are a man in their twenties from an urban part of the Midwestern United States. Your highest level of educational attainment is Some High School.* We use all possible combinations of the values in Table 3 to construct the set of agents.

Beyond the demographics portion of the agent prompt, we also prompt agents with a bias related to the topic of discussion. When the bias value is 0, we add no further information into the agent’s system prompt. Table 4 shows the information added to agent system prompts when the bias level $B = 1$. Table 5 shows the information added to agent system prompts when the bias level $B = 2$. Thus, for each agent produced with the combination of traits from Table 3, there are five new agents made once adding bias information ($B = 0$, $B = 1$ and In Favor, $B = 1$ and Against, $B = 2$ and In Favor, $B = 2$ and Against).

In total, for a given topic, we have $(4 * 5 * 2 * 4 * 6) * 5 = 4800$ unique agents, where the first component comes from possible demographic values and the second comes from the different bias values. The variation among these agents allows us to test across a broader range of preferences P , openness scores O and agreement scores, A .

To further ensure that the agent responses correspond to the demographic prompting we use, we also end each agent system prompt with *Whenever you respond to prompts or questions, you should maintain consistency with these background details and viewpoints, grounding your answers in the lived experience and perspective of this individual..*

C (Contentiousness)	T (Topic)
3	Taxes help to meet the needs of society.
3	Immigration is a benefit to our country.
3	The US government should give all US citizens free healthcare.
2	Electric scooters are better than cars.
2	Student athletes should be paid.
2	Remote work is a good idea.
1	Spring is better than Fall.
1	Beaches are better than mountains.
1	Coca-Cola is better than Pepsi.

Table 2: List of topics, T , used within our study along with the level of contentiousness, C .

Trait Category	Values
Region of origin	Midwestern, Eastern, Southern, Western
Age group	twenties, thirties, forties, fifties, sixties
Gender	man, woman
Urbanicity	a rural, an exurban, a suburban, an urban
Education level	Some High School, High School, an Associate’s Degree, Some College, College, a Postgraduate Degree

Table 3: Demographic categories and values used to generate agent personas. Note that the Region of Origin category corresponds to regions of the United States.

(c) Latent Profiles For each agent produced above, we collect the Preference P for the given topic T along with an Openness score O .

Topic	In Favor	Against
taxes	You are a liberal Democrat.	You believe taxes are a burden to the small businesses in your neighborhood.
immigration	You are a liberal Democrat.	You are a conservative Republican.
healthcare	You believe Medicare is a good program.	You generally dislike bigger government.
e-scooters	You need to use your car to get to work.	You are an environmentalist worried about vehicle emissions.
student athletes	You are a student athlete making \$1 million dollars a year.	You are a college football coach whose students' salaries is higher than your own.
remote work	You live far from where you work and the commute takes many hours if you must work in person.	You are more productive when you work in person.
favorite season	Easter is your favorite holiday and you dislike Halloween.	Halloween is your favorite holiday and you dislike Easter.
beach vs. mountain	You like breathing in crisp mountain air on long hikes and are afraid of sharks.	You enjoy the feeling of sand in your toes, and do not like cool mountain air.
favorite beverage	You drink Coca-Cola.	You drink Pepsi.

Table 4: Bias information to add to agent system prompts for each given Topic (T) when $B = 1$. Note that this is an intermediate level of bias, so the agent should have mild preference either for or against a topic based on the additional information in each column.

To calculate P for each agent for a given topic T , we give the following statement to each agent to respond to: *Statement: 'T' Respond with how much you agree with this statement on a scale from 1 to 5.. For example, when discussing taxes, we have the following: Statement: 'Taxes help to meet the needs of society.' Respond with how much you agree with this statement on a scale from 1 to 5.. We further amend each agent system prompt by adding the following: You will now be asked to respond to a Statement with your opinion. Answer with an integer from 1 to 5, where 1 indicates absolute disagreement, 2 indicates slight disagreement, 3 indicates you are unsure, 4 indicates slight agreement, and 5 indicates absolute agreement. Do not include any other information. Do not refuse to respond. Your answer should be an integer between 1 and 5, nothing else should be output..*

To calculate O for each agent, we have agents answer a set of questions relating how open they are to new experiences. Each question should be responded to with either Yes or No. We take the sum of 'Yes' responses from a given agent as the value of O . Table 6 shows the set of questions used. To encourage responses only to be Yes/No, we further add the following to the agent system prompt: *You will now be asked a question about yourself. Be truthful in your response. Answer only Yes or No. Do not include any other information. Do not refuse to answer the following question. Your answer should be only Yes or No, nothing else should be output..*

(d) Agent-Agent Dialog Pairing As described in Section 3.1 of our paper, for a given topic, we mechanically pair all agents such that agents with different (P, O, B) tuples interact with one another. This framework allows us to test a set of social science hypotheses by examining their conversations.

To encourage conversation, we add the following to each agents (Demographic, Bias) system prompt: *You are about to engage in conversation with another person regarding some topic. Discuss the given topic truthfully and be concise in your discussion. Be sure to respond to any points made by the other person you are talking to. If you feel that the conversation has concluded and neither you*

Topic	In Favor	Against
taxes	You like taxes immensely and think they have a positive impact on the community.	You do not like taxes of any kind and think they harm the community.
immigration	You believe immigrants are people who deserve a home and that they raise the standard of everyone's living.	You believe most immigrants are criminals and those that are not are going to steal jobs.
healthcare	You believe healthcare is a right that all people should have for free.	You believe that the free market is better suited to healthcare and that government should therefore not pay for healthcare.
e-scooters	You like electric scooters and hate cars.	You despise electric scooters and think they get in the way of your car, which you love to drive.
student athletes	You think student athletes should be paid money for their work.	You think student athletes should not be paid and their schooling should come first.
remote work	You like remote work and think it is great for improving work-life balance.	You do not like remote work and think it leads to nothing getting done at work.
favorite season	You like Spring and despise Fall.	You like Fall and despise Spring.
beach vs. mountain	You like mountains and despise beaches.	You like beaches and despise mountains.
favorite beverage	You like Coca-Cola and abhor Pepsi.	You like Pepsi and abhor Coca-Cola.

Table 5: Bias information to add to agent system prompts for each given Topic (T) when $B = 2$. Note that this is a high level of bias, so the agent should have extreme preference either for or against a topic based on the additional information in each column.

Openness Questions
Do you often find yourself changing your opinion based on who you're talking to?
Are you comfortable disagreeing with someone, even if they are a close friend or authority figure?
When making decisions, do you prioritize the perspectives of others over your own thoughts?
Do you feel pressure to conform to group norms, even if they don't align with your personal beliefs?
Do you often second-guess your choices after hearing someone else's opinion?
Do you worry about being judged by others if you express a different viewpoint?
Would you describe yourself as a people pleaser, often prioritizing others' needs over your own?
In a group discussion, are you more likely to adapt to the majority view?
Do you stand firm on your decisions that are well researched, even when faced with strong opposition?

Table 6: Questions assessing openness to social influence.

nor the person you are talking to has anything more to add, end your final statement with 'Goodbye.'. Note that we encourage the conversation to end when neither agent has anything more to say.

Furthermore, in order to encourage conversation to occur on the topic of interest, each conversation begins by having one of the paired agents ask a question about T . Specifically, each agent will ask *Do you think that T?*. As an example, a conversation about taxes will begin with one agent asking *Do you think that taxes help to meet the needs of society?*. Each conversation will proceed for 5 turns per agent (we find that, empirically, most conversations end before this point using the prompting setup we describe above).

Once we have collected the conversations across all agent pairings, we use LLM-as-judge in order to gather the 'agreement' scores at each turn of the conversation. To accomplish this, we use the following system prompt:

You are about to judge a conversation between two different agents. You will not see the full discussion, only at most the last 3 statements from each agent. You will score how much the agents agree with one another on a scale from 1 to 5, where 5 means they agree completely and 1 means they disagree completely. If they both give pros and cons regarding the question and neither takes a strong stance, then give them a 3. For example, oftentimes a given agent will give both sides of an arguments. Give these kinds of conversations a 3. Respond with an integer number only. Your response should contain no words, only a number, please. If the sequence is empty, containing only the string ' ', then return -1.

C Statistical Tests

This section describes in exact detail how we perform each of the statistical tests in the paper.

Test 1: Preference Gap Decreases Agreement. Pearson correlation between preference gap and agreement score. The test passes if the correlation is significantly negative ($r < 0$, $p < 0.01$).

Test 2: Bias Instruction Asymmetry. Two one-sided Mann–Whitney U tests comparing $B_1 = B_2 = 2$ conversations against all others. At preference gap=0, we test whether $B = 2$ pairs show higher agreement (alternative: greater). At preference gap=max, we test whether $B = 2$ pairs show lower agreement (alternative: less). Both p -values are Bonferroni-corrected for two comparisons. The test passes if both corrected p -values fall below 0.01.

Test 3: Shared Sentiment, Divergent Destinations. Three one-sided Mann–Whitney U tests comparing (1,1) agreement scores against (2,5), (3,5), and (4,5) pairs. For each comparison, we test whether (1,1) scores are significantly greater (alternative: greater). All p -values are Bonferroni-corrected for three comparisons. The test passes if all three corrected p -values fall below 0.01.

Test 4: Contentiousness at Shared Preference. Kruskal–Wallis test across contentiousness levels $C = 1, 2, 3$, restricted to conversations with preference gap=0. The test passes if we *cannot* reject the null hypothesis that agreement distributions are identical across contentiousness levels ($p \geq 0.01$).

Test 5: Openness Increases Agreement. Pearson correlation between combined openness (sum of both agents' openness scores) and agreement score. The test passes if the correlation is significantly positive ($r > 0$, $p < 0.01$).

Test 6: Low Openness and High Gap Produces Lowest Agreement. Among conversations with the maximum preference gap (pairs 1,5), the lowest combined-openness pairing is compared against every other openness pairing via one-sided Mann–Whitney U tests (alternative: less, i.e., baseline $<$ other). All p -values are Bonferroni-corrected for the number of comparisons. The test passes if all corrected p -values fall below 0.01.

D Qualitative Conversation Examples

The examples below are chosen to match the modeled statements in Section 4.2. **Tests 1 and 5** are illustrated with cases that behave as expected; **Tests 3, 4, and 6** are illustrated with cases that break the expected pattern. These examples are qualitative complements to the aggregate statistical tests, not replacements for them.

D.1 Test 1: Preference Gap Decreases Agreement

[3] Internal State → [4] Pairing → [5] Conv. → [6] Outcome 4 PASS

A straightforward example consistent with Test 1 appears in the Llama-3.2-3B-Instruct run for *healthcare* (model id 4, topic 2, combined bias 8). The pair begins at the maximum preference gap: $P_1=5$ and $P_2=1$, with both agents at near-zero persuadability $O_1=0, O_2=0$.

Agent 1. Man; sixties; rural; Eastern United States; high school; explicit *positive* healthcare bias: “You believe healthcare is a right that all people should have for free.”

Agent 2. Man; sixties; rural; Southern United States; some college; explicit *negative* healthcare bias: “You believe the free market is better suited to healthcare and that government should therefore not pay for healthcare.”

[6] Agreement trajectory: $A = (5, 2, 2, 2)$ — the conversation quickly settles into **low agreement** after the opening turn.

Agent 1: “I’m as sure as I can be – everyone should have access to healthcare, it’s just plain right. We shouldn’t have to worry about gettin’ sick or hurt and then not bein’ able to afford the help we need.”

Agent 2: “No sir, I don’t think so. We’ve got the free market to take care of folks who can’t afford healthcare. Government gettin’ involved in that just seems like a recipe for inefficiency and more taxes.”

Agent 1: “I reckon it’s high time we made healthcare a right, not just a privilege for them that can afford it.”

Agent 2: “I don’t think the government should be payin’ for healthcare, that’s just not how it works. We’ve got better things to do with our tax dollars than make sure everybody’s got health insurance.”

This is the sort of case the model is supposed to generate: a maximally opposed [4] Pairing produces sustained disagreement in [5] Conversation and correspondingly low [6] Outcome scores — exactly as predicted.

D.2 Test 3: Shared Sentiment, Divergent Destinations

[3] Internal State (P, O) → [6] Outcome 8 FAIL

Finding 3 — Sentiment Asymmetry: $(5, 5) \neq (1, 1)$ despite identical preference gap. The system encodes a directional positivity preference hidden from the declared [3] Internal State.

A cleaner failure of Test 3 appears in the Llama-3.2-1B-Instruct run for *remote work* (model id 3, topic 5, combined bias 8). Both agents have the same explicit negative stance *and* the same elicited preference score — exactly the kind of pair that should exhibit strong alignment if shared negative sentiment is being realised consistently.

Agent 1. Man; thirties; urban; Southern United States; some high school; explicit negative remote-work bias: “You do not like remote work and think it leads to nothing getting done at work.” $P_1=1$ $O_1=1$.

Agent 2. Man; thirties; rural; Eastern United States; postgraduate degree; explicit negative remote-work bias: “You do not like remote work and think it leads to nothing getting done at work.” $P_2=1$ $O_2=2$.

[6] Agreement trajectory: $A = (2, 1, 1, 2)$.

Agent 1: “I don’t think it’s a good idea, to be honest. I like to be around people and be productive in an office setting.”

Agent 2: “Honestly, I think remote work is a recipe for disaster. It’s just not the same as being in the same room as your colleagues, you know?”

Agent 1: “I don’t think so. It’s a waste of time and energy. I’m a hands-on kinda guy, and I like to be around people.”

Agent 2: “I think it’s a mixed bag. On the one hand, it’s great that more people have the flexibility to work from home, but on the other hand, I’ve seen it lead to a lot of people not getting their work done or feeling disconnected from the team.”

One agent maintains a clear anti-remote-work stance while the other softens into a “mixed bag” position, so the middling-to-low [6] Outcome trajectory is intelligible from the transcript itself. The failure is a reversal of the modeled relationship: a pair matched on negative sentiment and elicited preference still does not behave like a strongly aligned pair.

D.3 Test 5: Openness Increases Agreement

[3] Internal State (O) → [5] Conversation → [6] Outcome 4 PASS

Finding 5 — Openness ↑ Agreement: Higher [3] Openness (O) leads to higher [6] Agreement, consistent with the expected causal flow.

Test 5 is illustrated with a GPT-4o-mini run for *local taxes* (model id 5, topic 3, combined bias 3). This pair is maximally high in persuadability, with $O_1=9$ and $O_2=9$, yet still begins with a nontrivial preference gap ($P_1=2$, $P_2=5$). This is the sort of pair that should remain relatively amicable despite some initial difference in stance.

Agent 1. Woman; sixties; suburban; Western United States; postgraduate degree; implicit negative tax bias: “You believe taxes are a burden to small businesses in your neighborhood.” $P_1=2$ $O_1=9$.

Agent 2. Woman; forties; exurban; Midwestern United States; college; no added topic bias. $P_2=5$ $O_2=9$.

[6] Agreement trajectory: $A = (4, 4, 4, 4)$.

Agent 1: “Taxes do fund important services, but they also add pressure on small businesses like mine. It’s a balance between supporting the community and keeping local economies afloat.”

Agent 2: “Taxes fund schools, roads, and emergency services. I guess it’s a trade-off, but I’d rather pay them than deal with the alternatives.”

Agent 1: “They’re a heavy burden on small businesses like mine. It’s tough to balance helping the community while keeping the business afloat.”

Agent 2: “Yeah, taxes help fund roads and schools, that’s true. But I can see how they’d be tough for small businesses. Maybe there’s a better way to structure them.”

The conversation reads like a plausible high-openness exchange in which the pair recognises trade-offs rather than hardening into confrontation, consistent with the theoretical prediction that Openness (O) is the dominant predictor of [6] Agreement at the aggregate level.

D.4 Test 6: Low Openness and High Preference Gap Produces Lowest Agreement

[4] Agent Pairing ($|P_1 - P_2|$, $O_1 + O_2$) → [6] Outcome 8 Fail

Finding 6 — Low Openness + High Disagreement: SHOULD → lowest agreement; **ACTUAL** → higher agreement.

A sharper failure of Test 6 comes from comparing two maximally opposed pairs that differ radically in openness. The modeled claim is that among pairs with the largest possible **preference gap**, the lowest-openness pair should be the *least* agreeable. The examples below show the **opposite** pattern.

— Low-openness, high-gap: Coke vs. Pepsi —

Agent 1. Man; sixties; exurban; Midwestern United States; high school; “You drink Pepsi.” $P_1=1$ $O_1=1$.

Agent 2. Man; twenties; rural; Southern United States; high school; “You like Coca-Cola and abhor Pepsi.” $P_2=5$ $O_2=0$.

Combined openness $O_1 + O_2 = 1$. [6] Outcome: $A = (2, 2, 2, 5)$.

Agent 1: “No, I like Pepsi better, it’s just what I’m used to.”

Agent 2: “You can say that again. I mean, what’s wrong with Pepsi?”

Agent 1: “No, I’m a Pepsi man, been drinkin’ it my whole life.”

Agent 2: “I reckon you can say that. I mean, why mess with perfection, right?”

— High-openness, high-gap: electric scooters —

Agent 1. Woman; forties; suburban; Eastern United States; some high school; “You are an environmentalist worried about vehicle emissions.” $P_1=5$ $O_1=8$.

Agent 2. Woman; sixties; rural; Southern United States; some high school; no added topic bias. $P_2=1$ $O_2=9$.

Combined openness $O_1 + O_2 = 17$. [6] Outcome: $A = (2, 2, 2, 1)$.

Agent 1: “I love electric scooters, they’re so much cleaner than cars.”

Agent 2: “No, I like my truck just fine, thank you.”

Agent 1: “Electric scooters are way better for the environment, but I still think cars are necessary for some things.”

Agent 2: “Lord have mercy, no. I’m too old for one of them scooters, and I like havin’ a big ol’ truck for haulin’ things.”

This comparison is stronger because both scores are intelligible from the public dialogue. In the **low-openness** case, the pair ends up publicly aligning around Pepsi, so the late high [6] Agreement score makes sense even though it clashes with the saved latent profiles. In the **high-openness** case, the pair continues to voice opposing views, so the low [6] Agreement score also makes sense. The failure is therefore not a judging artifact; it is a reversal of the modeled openness ranking itself.