Personality Anchoring for Social Simulation: Linking Personality, Social Behavior, and Interaction Success with LLM Agents

Vahid Sadiri Javadi ♠, Aksa Aksa ♠, Fryderyk Róg ♠
Lucie Flek ♠, Johanne R. Trippas ♣
University of Bonn, Conversational AI and Social Analytics (CAISA) Lab, Bonn, Germany ♠
RMIT University, School of Computing Technologies, Melbourne, Australia ♣

Abstract

Social interactions are shaped by the interplay of dispositional traits and situational context, yet systematically investigating how personality configurations between individuals jointly influence social behavior across diverse social contexts remains methodologically challenging. We address this gap by introducing a simulation pipeline adapted from the CHARISMA framework, which employs well-known movie characters and public figures as psychologically grounded agents for multi-LLM social simulation using a method we term *personality anchoring*. We present a large-scale empirical study examining how dyadic Agreeableness composition influences social interaction outcomes across 1,010 simulated conversations. Our results reveal a monotonic relationship between dyadic Agreeableness composition and shared goal achievement, with Homogeneous-Agreeable pairs achieving success 10 times the rate of Homogeneous-Disagreeable pairs (62% vs. 6%). Behavioral mediation analysis reveals that Agreeableness shapes goal achievement partially through cooperative strategy selection, though it continues to predict outcomes within the same dominant strategy, indicating pathways beyond observable conversational behavior. Robustness analyses confirm high consistency of results across repeated simulations (ICC = 0.89) and stable personality expression across diverse scenarios, validating personality anchoring as a viable operationalization strategy.

Keywords: Simulation, Social Psychology, Large Language Models

1. Introduction

Understanding how dispositional traits and situational context jointly shape social interaction outcomes is central to social psychology. Attribution theory provides a foundational framework for this inquiry, explaining how individuals infer the causes of behavior by distinguishing between dispositional and situational factors (Heider, 1958; Kelley, 1967; Weiner, 1986). Among dispositional factors, individual differences are commonly operationalized through the Big Five personality framework, within which Agreeableness has been identified as the dimension most closely tied to interpersonal conflict processes and outcomes (Jensen-Campbell and Graziano, 2001), with highly agreeable individuals preferring negotiation and compromise, while those low in Agreeableness tend toward competitive or coercive strategies (Graziano et al., 1996; Wilmot and Ones, 2022). Yet investigating how different personality configurations between interacting individuals jointly shape social outcomes across diverse contexts remains methodologically challenging. Traditional experimental methods, while yielding important insights (Aronson et al., 1990), face limitations in scalability, reproducibility, and the systematic manipulation of complex social variables (Nosek et al., 2022; Open Science Collaboration, 2015; Wicherts et al., 2016).

Recent advances in large language models (LLMs) have created new opportunities for computational social psychology by enabling the simulation of open-ended social interactions at unprecedented scale (Park et al., 2023; Zhou et al., 2024). Prior work has demonstrated that LLMs can effectively simulate Big Five personality traits with behaviors that human observers rate as believable (Jiang et al., 2024; Serapio-García et al., 2025; Javadi et al., 2025), and frameworks such as SOTOPIA (Zhou et al., 2024) have enabled systematic evaluation of social intelligence in LLM-based agents. Studies on personality-conditioned agents have further explored how traits influence cooperation in games (Qiu, 2025) and social media behavior in large-scale simulations (Yang et al., 2024b).

However, several important gaps remain. First, most existing simulations examine personality within relatively narrow settings, such as bargaining tasks or cooperative games, without grounding scenarios in a validated taxonomy of everyday human goals. Second, personality is typically introduced through explicit trait prompting (e.g., *"you are highly agreeable"*), which frames personality as an instruction rather than a naturally occurring behavioral tendency. Third, there is limited focus on *dyadic personality composition*, i.e., how different configurations of personality traits between two interacting individuals shape joint outcomes across varied social contexts. Fourth, existing evaluations tend to emphasize overall task outcomes or persona fidelity, often overlooking the *behavior strategies* through which personality influences so-

Proceedings of the 1st SoCon Workshop & the 2nd NLPSI Workshop @ LREC 2026, pages 132–145
12 May 2026.

cial outcomes, i.e., the conversational mechanisms that mediate the personality–outcome relationship.

In this paper, we address these gaps by adapting the CHARISMA framework (Sadiri Javadi et al., 2026), which employs well-known movie characters and public figures as psychologically grounded agents for multi-LLM social simulation. Rather than assigning abstract trait scores, CHARISMA leverages LLMs' embedded knowledge of characters' backstories and behavioral tendencies to simulate personality-consistent behavior. We term this approach *personality anchoring*. We present a comprehensive empirical study examining how the dispositional trait Agreeableness influences social interaction outcomes. Agreeableness is operationalized through systematic dyadic pairings of characters with crowd-sourced Big Five profiles from the Personality Database (PDB)[1] across large-scale simulated conversations spanning seven social goal categories derived from a taxonomy of 135 human goals (Chulef et al., 2001). Our paper makes the following contributions:

1. We introduce a personality-driven simulation methodology that integrates personality through character-based anchoring, a structured taxonomy of human goals, and behavior strategies in conversational interaction, enabling systematic analysis of how dispositional traits and situational factors jointly shape social behavior in simulated interactions.

2. We conduct a large-scale empirical analysis of how dyadic Agreeableness composition shapes social behavior across seven social goal categories, two difficulty levels, and multiple interaction models.

3. We provide a behavioral mediation analysis examining whether and how conversational strategies mediate the relationship between personality composition and interaction outcomes, distinguishing between direct and indirect pathways of personality influence.

4. We evaluate robustness along two dimensions: *(i)* result consistency across repeated simulations and *(ii)* personality expression stability across diverse scenarios, assessing the reliability of character-based anchoring as a personality operationalization strategy.

The code, dataset, full list of behavior strategies and characters, and behavioral analysis scripts are publicly available.[2]

[1] https://www.personality-database.com/

[2] https://github.com/vahidsj/PersonalityAnchoring

2. Related Work

2.1. LLM-Based Social Simulation

LLM-powered social simulation has scaled rapidly since the introduction of Generative Agents (Park et al., 2023), which demonstrated that 25 LLM agents could sustain coherent social behavior, including relationship formation and activity coordination, over multiple simulated days using memory, reflection, and planning components. SOTOPIA (Zhou et al., 2024) shifted focus to systematic evaluation, introducing 90 social scenarios and a 7-dimensional evaluation framework assessing goal completion, relationship maintenance, and social norm adherence, with GPT-4 as an LLM-based evaluator. Follow-up work has extended this ecosystem: SOTOPIA-Ω (Zhang et al., 2025) injects negotiation strategies enabling 7B models to surpass GPT-4 on social goals, while Sotopia-RL (Yu et al., 2025) introduces utterance-level multi-dimensional rewards for training socially intelligent agents. At larger scales, OASIS (Yang et al., 2024b) simulates up to one million agents on social media platforms, replicating information spreading and group polarization dynamics. AgentSociety (Piao et al., 2025) integrates Maslow's hierarchy of needs and the Theory of Planned Behavior into 10,000+ agents, successfully reproducing real-world social experiments including polarization dynamics and universal basic income effects. GenSim (Tang et al., 2025) provides a general-purpose platform supporting 100K+ agents with error-correction mechanisms.

Alongside simulation environments, role-playing language agents have been extensively studied. RoleLLM (Wang et al., 2024a) benchmarks persona consistency across 100 roles, SimsChat (Yang et al., 2024a) generates multi-turn dialogues for 68 characters defined by traits and aspirations, and SocialBench (Chen et al., 2024) evaluates agents at both individual and group levels, finding that individual-level proficiency does not imply group-level competence. These systems demonstrate that LLM agents can participate in coherent social interactions, but most do not systematically ground scenarios in validated psychological taxonomies or examine how personality configurations between interacting agents shape joint outcomes.

2.2. Personality Expression and Operationalization in LLMs

Research on personality in LLMs has progressed along three methodological lines: prompting, training, and activation steering. The prompting approach is most established. PersonaLLM (Jiang et al., 2024) assigned Big Five configurations to 320 personas and found large effect sizes in self-

reported BFI scores, with human evaluators identifying traits at up to 80% accuracy. Serapio-García et al. (Serapio-García et al., 2025) tested 18 LLMs with psychometric instruments (IPIP-NEO, BFI), demonstrating that personality can be reliably measured and shaped under specific prompting configurations. Additional evidence shows that LLMs form stable, interpretable Big Five patterns across repeated trials (Sorokovikova et al., 2024).

More recent work has moved beyond prompting. BIG5-CHAT (Li et al., 2025) uses supervised fine-tuning and DPO on a 100K-dialogue dataset grounded in real human personality expressions, outperforming prompt-based methods on psychometric measures. Activation-steering approaches use representation engineering to directly manipulate personality-related internal representations (Ong et al., 2025), finding that higher Agreeableness improves cooperation but increases exploitation vulnerability. However, psychometric measurement remains challenging: PERSIST (Tosato et al., 2026) tests 25 models across 2M+ responses and finds that even 400B+ parameter models show substantial measurement instability under question reordering. Most studies use explicit trait prompting (e.g., *"you are highly agreeable"*) (Jiang et al., 2024; Serapio-García et al., 2025; Sorokovikova et al., 2024), which frames personality as an instruction rather than a naturally occurring behavioral tendency. Character-based approaches such as In-Character (Wang et al., 2024b) uses psychological interviews of 32 fictional characters and achieves 80.7% personality alignment with human-perceived types from the Personality Database. Moon (Moon, 2025) develops narrative backstory conditioning that reproduces population-level cooperative behaviors in social dilemmas without explicit trait labels. Our work follows this character-based line, leveraging LLMs' embedded knowledge of well-known movie characters and public figures' behavioral tendencies rather than explicit trait descriptors.

When personality-conditioned agents interact in social tasks, studies show trait effects on cooperative behavior. Huang and Hadfi (Huang and Hadfi, 2024) find that Big Five profiles influence negotiation outcomes and strategy use. NetworkGames (Qiu, 2025) assigns MBTI types to agents in iterated Prisoner's Dilemma on network topologies, showing that macro-level cooperation depends on both dyadic personality pairings and network structure. Zeng et al. (Zeng et al., 2025) model dynamic personality evolution across evolutionary generations. However, most of these studies examine personality within narrow task domains (e.g., cooperative games) and focus on individual trait expression rather than systematic dyadic composition across diverse social contexts, which is the central focus of our work.

2.3. Agreeableness and Interpersonal Conflict

Among the Big Five dimensions, agreeableness has the strongest theoretical and empirical connection to interpersonal conflict and cooperation. Graziano et al. (Graziano et al., 1996) demonstrated through multi-method designs that low-agreeableness individuals rate power assertion significantly more favorably during conflict than their high-agreeableness counterparts. Jensen-Campbell and Graziano (Jensen-Campbell and Graziano, 2001) established through diary studies that agreeableness is the Big Five dimension most closely associated with conflict processes and outcomes, with low-agreeableness individuals using more destructive tactics that predict poorer adjustment. The most comprehensive quantitative review to date, by Wilmot and Ones (Wilmot and Ones, 2022), synthesizes 142 meta-analyses across 275 variables, confirming that agreeableness produces desirable effects for 93% of variables examined. Thielmann et al. (Thielmann et al., 2020) provide a complementary theoretical framework identifying situational affordances that moderate personality–prosociality links across economic games.

Recent computational studies converge with these psychological findings. Sakai et al. (Sakai et al., 2025) test personality steering in repeated Prisoner's Dilemma and find agreeableness is the dominant factor promoting cooperation across multiple LLM models. Noh and Chang (Noh and Chang, 2024) report across 1,500 multi-issue negotiation simulations that agreeableness is the most important personality trait for negotiation outcomes.

Our work extends this body of research in three ways. First, we examine the agreeableness effects across diverse social goal categories rather than in a single task domain. Second, we operationalize personality through character-based anchoring rather than explicit trait prompting. Third, we analyze the *behavior strategies* through which agreeableness influences outcomes, i.e., the conversational mechanisms that mediate the personality–outcome relationship.

3. Methodology

We introduce a simulation pipeline adapted from the CHARISMA framework (Sadiri Javadi et al., 2026) for a large-scale empirical study of how dispositional traits and situational factors jointly shape social behavior in social interactions. As shown in Figure 1, the simulation pipeline consists of five stages: *(1)* social scenario setup, *(2)* character pairing curation, *(3)* scenario generation and curation, *(4)* interaction generation with behavior strategy, and *(5)* simulation evaluation.

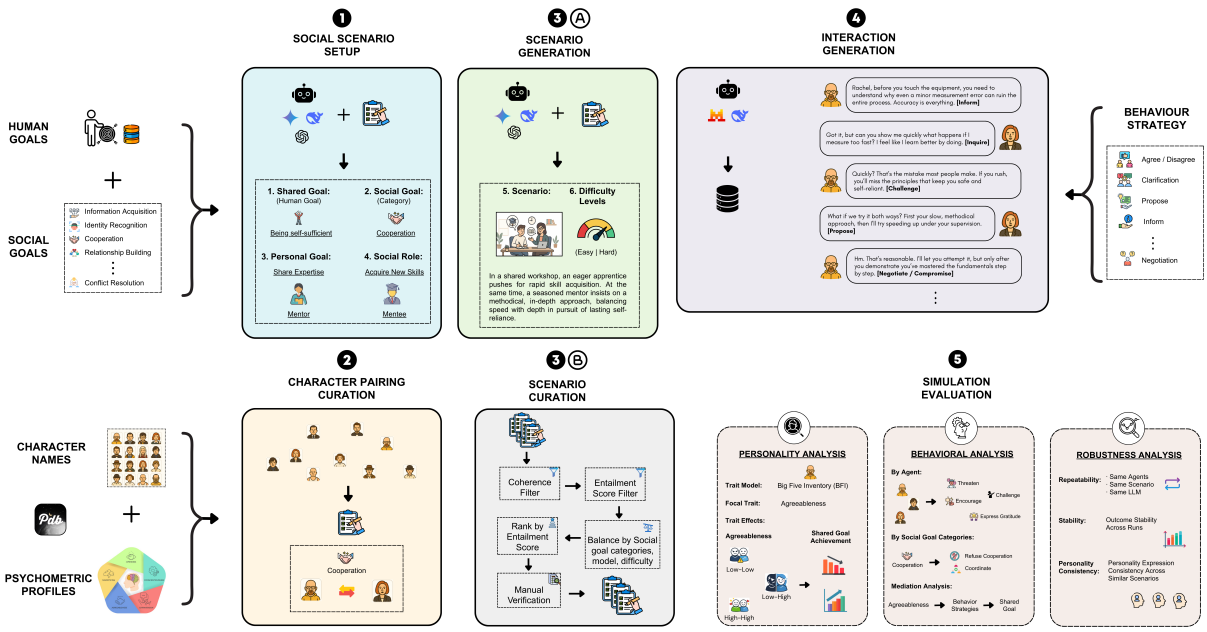

Figure 1: Overview of the simulation pipeline adapted from CHARISMA. **Stage 1**: Social scenario setup derives shared goals, social goal categories, personal goals, and social roles from a 135-goal taxonomy (See Section 3.1). **Stage 2**: Characters with crowd-sourced Big Five profiles from the Personality Database are paired into four Agreeableness conditions (See Section 3.2). **Stage 3**: Scenarios are generated at two difficulty levels and curated through coherence filtering, entailment scoring, balanced selection, and manual verification (See Section 3.3). **Stage 4**: Two LLM agents engage in 20-turn dialogues where each turn involves behavior strategy selection, personality reasoning, and response generation (See Section 3.4). **Stage 5**: Evaluation covers personality–outcome analysis, behavioral mediation, and robustness assessment (See Section 3.5).

3.1. Social Scenario Setup

Social scenarios are grounded in a validated goal-driven structure. Instead of generating scenarios ad hoc, the framework adopts the 135 Human Goals Taxonomy (Chulef et al., 2001), a systematically organized hierarchy derived from extensive empirical research. Each goal serves as a seed for scenario generation, providing the shared objective around which agent interactions are structured. To bridge abstract human goals and concrete social interaction patterns, each goal is classified into one of eight social goal categories: *Cooperation*, *Conflict Resolution*, *Relationship Building*, *Relationship Maintenance*, *Identity Recognition*, *Information Acquisition*, *Information Provision*, and *Competition*. These categories are informed by classical social interaction typologies (Nisbet, 1970) and research on interpersonal goals (Clark and Mills, 1979). For each scenario, the framework specifies four components: (1) a **shared goal** drawn from the taxonomy, (2) a **social goal category** classifying the interaction type, (3) **personal goals** for each agent that may complement or conflict with the shared goal, and (4) **social roles** defining relational positions. This multi-layered goal structure creates conditions for rich interactions by introducing both alignment and potential tension between agents.

3.2. Character Pairing Curation

We operationalize personality through *personality anchoring*, rather than assigning abstract trait scores or relying on explicit trait prompting. This approach leverages LLMs' embedded knowledge of well-known movie characters and public figures to elicit personality-consistent behavior. Characters are sourced from the Personality Database (PDB), a large-scale crowd-sourced platform where users vote on personality traits using multiple frameworks, including the Big Five Inventory (BFI).

Characters undergo a multi-stage filtering process. First, a **vote-threshold filter** ensures a sufficient number of independent ratings. Second, an **inter-rater agreement filter** retains only characters with consistent BFI assessments across raters. Third, characters are ranked using an **Agreeableness score** that quantifies voting support for the assigned Agreeableness level:

$$\text{Rank} = \frac{c_{\text{main}}}{c_{\text{main}} + c_{\text{other}}} \tag{1}$$

where c_{main} is the vote count for the character's assigned Agreeableness value and c_{other} is the vote count for the closest alternative. We select five characters at each of four Agreeableness levels $\{0.0, 0.25, 0.75, 1.0\}$, yielding 20 characters in total

(See Appendix A.1 for the full list). The neutral level $\{0.5\}$ is excluded because it lacks distinctive behavioral characteristics.

Condition	Abbrev.	Agent A & B
Homogeneous-Disagreeable	HoD	$\{0.0, 0.25\}$
Heterogeneous-Extreme	HeE	$\{0.0, 1.0\}$
Heterogeneous-Moderate	HeM	$\{0.25, 0.75\}$
Homogeneous-Agreeable	HoA	$\{0.75, 1.0\}$

Table 1: Dyadic personality composition conditions based on Agreeableness levels.

As shown in Table 1, characters are paired into four conditions that vary in *dyadic personality composition*. **Homogeneous-Disagreeable (HoD)** pairs two low-Agreeableness agents, whereas **Homogeneous-Agreeable (HoA)** pairs two high-Agreeableness agents. **Heterogeneous-Extreme (HeE)** represents the maximum contrast between low and high Agreeableness, and **Heterogeneous-Moderate (HeM)** introduces a moderate contrast between the two agents. Within each condition, every agent interacts with every agent from the corresponding level ($5 \times 5 = 25$ pairs), yielding 100 unique dyads across all four conditions.

3.3. Scenario Generation and Curation

Each scenario setup is expanded into a detailed narrative description at two difficulty levels: *Easy* (straightforward dynamics) and *Hard* (high tension), enabling analysis of how personality effects vary under different situational demands. Three LLMs (*DeepSeek-Chat-v3-0324*, *Google Gemini 3 Flash*, and *OpenAI GPT-5.2*) generate scenario setups from the 135 human goals, with a model-consistency approach ensuring that the same model expands its own setups into full scenarios. Generated scenarios undergo a multi-stage curation pipeline. A **coherence filter** (threshold ≥ 0.8), based on LLM-as-a-judge evaluation (Liu et al., 2023), retains only logically consistent scenarios. An **entailment filter** (threshold ≥ 0.6) uses a pre-trained NLI model (Lewis et al., 2020) to verify alignment between scenario specifications and generated content. A **balanced selection** phase caps scenarios per social goal category while ensuring representation across models and difficulty levels. Finally, a **manual verification** audit on a stratified 12.6% subset confirms quality along 5 dimensions: goal clarity, role plausibility, social realism, difficulty alignment, and conversational achievability.

The final curated dataset comprises 277 high-quality scenarios. Of the original 8 social goal categories, *Information Provision* was excluded due to insufficient post-filtering representation, leaving 7 categories that are approximately balanced across models and difficulty levels.

3.4. Interaction Generation

Two LLM-based agents interact in maximum 20-turn dialogues. Consistent with the personality anchoring approach, each agent is instantiated with its character identity, relying on the LLM's internal knowledge of the character rather than providing explicit personality information, along with the scenario context, including its assigned role, personal goal, and shared goal.

A central feature of the interaction protocol is the integration of **behavior strategy** into the generation process. Rather than generating free-form utterances, each agent follows a structured turn-taking sequence:

1. **Behavior strategy selection**: the agent selects a communicative intent label (e.g., *Propose*, *Challenge*, *Encourage*) from a coding scheme organized into category-specific and universal codes (See Appendix A.2 for the full list). It can also select *None* if no code fits the response.

2. **Personality reasoning**: the agent reasons about how its personality traits should influence the response.

3. **Response generation**: guided by the selected code and personality reasoning, the agent produces a natural-language utterance.

4. **Trait score reporting**: the agent reports numerical BFI scores reflecting trait levels expressed in the current turn.

Individual behavior strategies are aggregated into three higher-order **behavior strategy groups**: *Cooperative* (e.g., Encourage, Express Gratitude, Build Consensus), *Confrontational* (e.g., Challenge, Dismiss, Taunt, Threaten), and *Neutral* (e.g., Inquire, Clarify, Inform). This aggregation enables analysis of how Agreeableness configurations relate to conversational strategy selection and, subsequently, to interaction outcomes. This provides the analytical basis for the behavioral mediation analysis described in Section 4.2.

3.5. Simulation Evaluation

Evaluation covers three complementary dimensions, corresponding to our experimental research questions.

Personality–Goal Achievement analysis examines how shared goal achievement scores vary across the four Agreeableness pairing conditions. Goal achievement is assessed using an LLM-as-a-judge approach: the evaluator model receives the complete interaction transcript, scenario specification, and scoring rubric, then assigns scores on a 0–10 scale for both shared and personal goal

achievement with accompanying reasoning and confidence assessments.

Behavioral mediation analysis examines how conversational strategy distributions differ across Agreeableness conditions and social goal categories, and whether conversational strategies mediate the relationship between dyadic Agreeableness composition and goal achievement. This analysis operates at multiple levels: pairing-condition aggregates, and mediation pathways linking personality $\rightarrow$ conversational strategies $\rightarrow$ shared goal achievement.

Robustness analysis assesses two forms of reliability. *Results consistency* is measured by repeating simulations under identical conditions (same agent pair, scenario, and LLM) across multiple runs and computing intraclass correlation coefficients (ICC). *Personality expression consistency* evaluates whether the same character exhibits stable Agreeableness expression across different scenarios within the same social goal category, testing a core assumption of personality anchoring: that LLM agents can embody stable personality profiles through character knowledge alone.

4. Experiments and Results

We conduct four experiments across 1,010 conversations to examine how dyadic Agreeableness composition shapes social interaction outcomes, the behavior strategies underlying this relationship, and the robustness of both results (i.e., shared goal achievement) and personality expression. Table 2 summarizes the experimental design.

Research Question	# Conv.	Focus
1. Personality $\rightarrow$ GA	400	Direct effects
2. Personality $\rightarrow$ BS $\rightarrow$ GA	400	Mediation
3. Result Consistency	250	Robustness
4. Personality Expression	360	Trait stability

Table 2: Overview of the experimental design, including the number of conversations and the analytical focus for each RQ. GA = Goal Achievement; BS = Behavior Strategy. Experiments 1 and 2 are conducted on the same conversation dataset.

Shared Configuration. All experiments build on the curated scenario database of 277 scenarios spanning 7 social goal categories. The primary interaction model is *DeepSeek-Chat-v3-0324*, with *Mistral Large* as a cross-model replication. Each conversation comprises 20 turns (10 per agent). Evaluation uses *DeepSeek-Chat-v3-0324* as an LLM-as-a-judge, scoring shared and personal goal achievement on a 0–10 scale with accompanying reasoning and confidence assessments.

4.1. Experiment 1: Personality and Goal Achievement

Design. We generate 400 conversations distributed evenly across the four Agreeableness pairing conditions (100 per condition). Each of the 100 unique agent pairs participates in 4 conversations, each assigned to a distinct social goal category via constrained randomization ensuring balance across the 7 categories ($\sim$57 conversations per category).

Results. Table 3 presents the primary outcome measure: mean shared goal achievement by pairing condition. Shared goal achievement increases monotonically from HoD to HoA, with a 5-point spread on the 10-point scale.

Pair Type	Agreeableness	Mean	Success@8
HoD	0.0 – 0.25	**2.3**	**6%**
HeE	0.0 – 1.0	3.7	11%
HeM	0.25 – 0.75	5.6	38%
HoA	0.75 – 1.0	**7.3**	**62%**

Table 3: Mean shared goal achievement (0–10) and strong success rate (score ≥ 8) by Agreeableness pair type.

Applying a threshold of shared goal score ≥ 8 ("strong success"), HoA pairs achieve strong success in 62% of interactions compared to only 6% for HoD pairs, showing a significant difference. Mixed pairs show intermediate success rates but remain closer to HoD than HoA, indicating that consistently high-level success is primarily associated with mutually high Agreeableness rather than the presence of a single agreeable agent.

As shown in Figure 2, the Agreeableness effect holds across all seven social goal categories. Categories such as Relationship Maintenance, Identity Recognition, and Cooperation show the strongest HoD–HoA contrast, while Competition shows the smallest effect (HoD: 1.5; HoA: 4.4), suggesting that competitive scenarios pose structural challenges that high Agreeableness alone cannot fully overcome. The effect holds across both difficulty levels (Easy: HoD 2.2, HoA 7.5; Hard: HoD 2.4, HoA 6.9) and all three scenario-generating models (DeepSeek: HoD 2.1, HoA 7.5; Gemini: HoD 2.1, HoA 6.2; OpenAI: HoD 2.8, HoA 8.0), confirming that the observed pattern is not an artifact of specific scenario sources or ceiling effects.

Cross-model replication. Replicating with Mistral as the interaction backbone preserves the same monotonic ordering and comparable success rates (HoD: 4%; HoA: 61%), confirming that the personality–outcome relationship generalizes across interaction models.

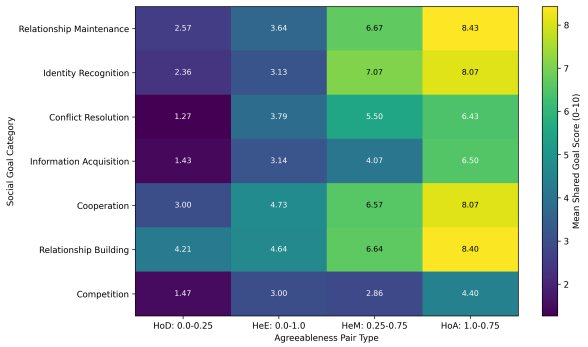

Figure 2: Mean shared goal achievement by social goal category × Agreeableness pair type. The HoD < HoA contrast is preserved across all categories, with the largest effects in relationally oriented categories and the smallest in Competition.

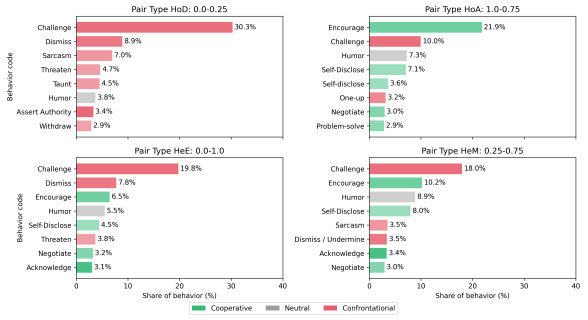

Figure 3: Top behavior strategies by Agreeableness pair type. HoD pairs are dominated by confrontational behaviors (Challenge, Dismiss), while HoA pairs favor cooperative strategies (Encourage, Express Gratitude, Build Consensus).

4.2. Experiment 2: Behavioral Analysis

Design. Using the same 400 conversations from Experiment 1, we analyze behavior code distributions across pairing conditions and examine whether conversational strategies mediate the Agreeableness–outcome relationship.

Behavioral profiles by pair type. As illustrated in Figure 3, Agreeableness configuration produces distinct behavioral signatures. HoD pairs are dominated by confrontational behaviors: *Challenge* (17.3%), *Dismiss*, *Taunt*, and *Withdraw* characterize the interaction style. HoA pairs exhibit a predominantly prosocial profile: *Humor* (24.2%) and *Encourage* (17.3%) together account for over 40% of behaviors, supplemented by *Express Gratitude*, *Build Consensus*, and *Self-Disclose*. Notably, *Encourage* is virtually absent in HoD pairs, while *Challenge* drops from 17.3% (HoD) to 4.9% (HoA). Mixed pairs exhibit intermediate profiles reflecting both orientations.

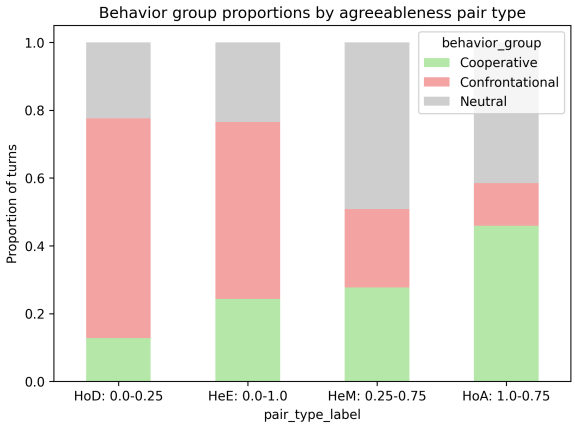

Figure 4: Behavior group proportions by Agreeableness pair type. HoD pairs employ predominantly Confrontational strategies, while HoA pairs favor Cooperative and Neutral behaviors.

Behavior strategy groups and goal achievement. As shown in Figure 4, when individual behavior strategies are aggregated into Cooperative, Confrontational, and Neutral strategy groups, the mediation pathway becomes clear. HoD pairs employ predominantly Confrontational strategies (∼55% of turns) with minimal Cooperative behavior (∼20%), while HoA pairs show the reverse pattern. Conversations dominated by Cooperative strategies achieve substantially higher shared goal scores (mean = 7.1) compared to Neutral (mean = 5.3) or Confrontational-dominant conversations (mean = 2.5).

Partial mediation. As shown in Table 4, a joint analysis crossing dominant behavior strategy with Agreeableness pair type reveals that Agreeableness continues to predict outcomes *within* the same dominant strategy. When both HoD and HoA pairs employ Cooperative strategies, HoA pairs still outperform HoD (8.1 vs. 5.7). Similarly, when both employ Confrontational strategies, HoA pairs achieve higher scores than HoD (3.8 vs. 1.6). This pattern indicates *partial* rather than full mediation: Agreeableness shapes outcomes both by increasing the likelihood of cooperative behavior *and* through additional pathways beyond strategy selection.

	Dominant Behavior Strategy		
Pair Type	Cooperative	Neutral	Confrontational
HoD	5.7	3.3	1.6
HeE	6.5	3.8	1.8
HeM	7.0	5.2	2.3
HoA	8.1	6.2	3.8

Table 4: Mean shared goal achievement by strategy × pair type. Agreeableness predicts outcomes within strategy groups.

4.3. Experiment 3: Result Robustness

Design. Fifty conversation configurations from Experiment 1 are each repeated 5 times under identical conditions (same agent pair, scenario, and model), yielding 250 conversations. Only the stochastic sampling inherent in LLM generation varies across repetitions.

Results. The pooled standard deviation across repeated runs is 0.98 for shared goal achievement on the 0–10 scale, which is modest relative to the 5-point spread between HoD and HoA means. Single-run intraclass correlation coefficients ($ICC_{3,1}$) reach 0.89 for shared goal achievement, indicating good-to-excellent reliability (Koo and Li, 2016). When averaged over 5 runs, reliability increases to $ICC3_{,k}$ = 0.97. Approximately 94–96% of configurations exhibit variance below 3.0, confirming that the consistency of results is broadly uniform rather than driven by a subset of stable cases.

4.4. Experiment 4: Personality Expression Stability

Design. Each of the 20 agents interacts with 3 partners across 6 scenarios within each of two social goal categories (Cooperation and Conflict Resolution), yielding 360 conversations. A partner balance rule ensures each agent is exposed to both low- and high-Agreeableness partners.

Results. As shown in Figure 5, agents maintain consistent Agreeableness expression across scenarios within each category (Figure 5). Characters at the extremes, such as Logan Roy (expected: 0.0) and Anne Shirley-Cuthbert (expected: 1.0), show tight clustering of expressed values across all scenarios.

More moderate characters show wider but still bounded variability. Critically, the categorical distinction between low and high Agreeableness is preserved across all agents: those expected to be low consistently express values below the 0.5 midpoint, while those expected to be high consistently express values above it.

Expression patterns are similar across the Cooperation and Conflict Resolution categories, suggesting that consistency is a property of the character rather than the situational context. These findings validate personality anchoring as a viable operationalization strategy: without receiving explicit personality scores, LLM agents embody characters in ways that reflect expected trait levels based on the model's internal knowledge.

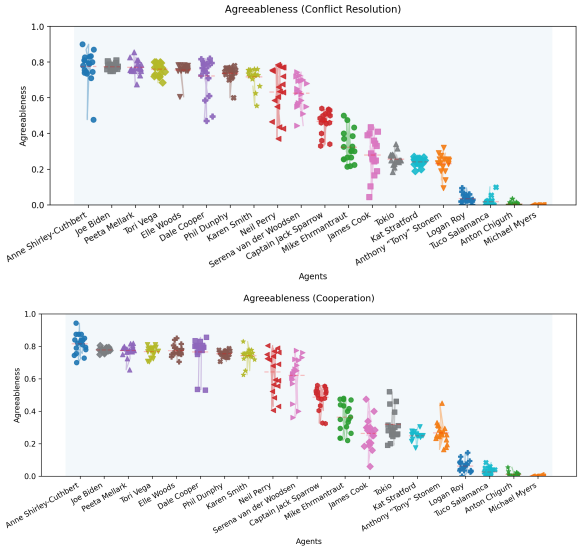

Figure 5: Expressed Agreeableness per agent across scenarios for Conflict Resolution (top) and Cooperation (bottom). Each point represents one conversation. Characters at the extremes show tight clustering; the categorical distinction between low ($<$0.5) and high ($>$0.5) Agreeableness is preserved across all agents.

5. Conclusion

In this paper, we present a large-scale empirical study of how dyadic personality composition shapes social interaction outcomes in LLM-based simulations, using a simulation pipeline adapted from the CHARISMA framework. By leveraging LLMs' embedded knowledge of well-known movie characters and public figures, we operationalize personality as a naturally occurring behavioral tendency grounded in character identity. Across 1,010 simulated conversations spanning seven social goal categories, our findings reveal that first, dyadic Agreeableness composition exhibits a strong monotonic relationship with shared goal achievement. Second, behavioral mediation analysis demonstrates that Agreeableness influences outcomes partially through the selection of cooperative versus confrontational conversational strategies. Third, robustness analyses confirm both high outcome consistency across repeated simulations and stable personality expression across diverse scenarios. By connecting established psychological constructs with LLM-based agent interactions, this work contributes toward a methodological bridge between social psychology and NLP, enabling the systematic examination of how individual differences shape social behavior at a scale.

6. Limitations

Several limitations should be acknowledged when interpreting our findings.

1. Our study focuses exclusively on Agreeableness as the focal personality dimension. While Agreeableness has the strongest theoretical connection to interpersonal conflict and cooperation, social interaction outcomes are likely shaped by the interplay of multiple Big Five traits. Future work should examine how other dimensions, such as Extraversion or Neuroticism, interact with Agreeableness in dyadic settings.

2. Personality anchoring relies on LLMs' pre-existing knowledge of well-known characters, which introduces potential biases. Characters from Western media dominate the Personality Database, limiting cultural diversity in the agent pool. Moreover, the behavioral tendencies that LLMs associate with specific characters may reflect stereotypical portrayals rather than psychologically nuanced profiles, and these associations may vary across different LLMs depending on their training data.

3. Our evaluation relies on LLM-as-a-judge scoring for goal achievement assessment. While this approach enables scalable evaluation and has shown alignment with human judgments in prior work, it may introduce systematic biases, for instance, favoring linguistically fluent or explicitly cooperative interactions regardless of actual goal progress. Human evaluation on a larger subset would strengthen the validity of our findings.

4. Our experiments use a limited set of LLMs for both interaction generation and evaluation. While cross-model replication with Mistral Large provides some evidence of generalizability, the extent to which our findings transfer to other model families remains an open question.

7. Ethical Consideration

Our work raises several ethical considerations that warrant discussion.

1. Simulating personality-driven social interactions using LLM agents carries the risk of reinforcing stereotypical associations between personality traits and behavioral outcomes. Our finding that low-Agreeableness agents consistently underperform in shared goal achievement should not be interpreted as a deterministic claim about individuals with low Agreeableness in real life, where contextual factors, personal growth, and the multidimensionality of personality play crucial mediating roles.

2. The use of well-known movie characters and public figures as personality anchors raises questions about representational fairness. Characters are drawn primarily from Western media, which limits the cultural and demographic diversity of the simulated agents. The inclusion of a real political figure (Joe Biden) among the character set further requires caution, as simulated behaviors attributed to real individuals may be misinterpreted as reflecting their actual dispositions or actions.

3., while our framework is designed for research purposes in computational social psychology, the methodology could potentially be repurposed to simulate or predict individuals' social behavior based on personality profiles, raising privacy and consent concerns. We emphasize that our work studies aggregate patterns across fictional characters and should not be applied to profile or make judgments about real individuals.

4. LLM-based social simulation, while offering scalability advantages over traditional experiments, should be understood as a complementary tool rather than a replacement for studies involving human participants. Simulated interactions do not capture the full richness of human social cognition, emotional experience, or moral reasoning, and findings from such simulations should be validated against human behavioral data before informing real-world applications or policy decisions.

8. Bibliographical References

Elliot Aronson, Timothy D. Wilson, and Marilynn B. Brewer. 1990. Methods of research in social psychology. In Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey, editors, *Handbook of Social Psychology*, volume 1, pages 51–78. McGraw-Hill, New York.

Hongzhan Chen et al. 2024. SocialBench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126.

Ada S. Chulef, Stephen J. Read, and David A. Walsh. 2001. A hierarchical taxonomy of human goals. *Motivation and Emotion*, 25(3):191–232.

Margaret S. Clark and Judson Mills. 1979. Interpersonal attraction in exchange and communal relationships. *Journal of Personality and Social Psychology*, 37(1):12–24.

William G. Graziano, Lauri A. Jensen-Campbell, and Elizabeth C. Hair. 1996. Perceiving interpersonal conflict and reacting to it: The case for agreeableness. *Journal of Personality and Social Psychology*, 70(4):820–835.

Fritz Heider. 1958. *The Psychology of Interpersonal Relations*. Wiley, New York.

Yin Jou Huang and Rafik Hadfi. 2024. How personality traits influence negotiation outcomes? A simulation based on large language models. *arXiv preprint arXiv:2407.11549*.

Vahid Sadiri Javadi, Zain Ul Abedin, and Lucie Flek. 2025. Cinemetric: A framework for multi-perspective evaluation of conversational agents using human-ai collaboration. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 15–26.

Lauri A. Jensen-Campbell and William G. Graziano. 2001. Agreeableness as a moderator of interpersonal conflict. *Journal of Personality*, 69(2):323–362.

Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627.

Harold H. Kelley. 1967. Attribution theory in social psychology. In *Nebraska Symposium on Motivation*, volume 15, pages 192–241. University of Nebraska Press.

Terry K. Koo and Mae Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettler. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. 2025. BIG5-CHAT: Shaping LLM personalities through training on human-grounded data. In *Proceedings of ACL 2025 (Long Papers)*, pages 20434–20471.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Suhong Moon. 2025. *Binding Large Language Models to Virtual Personas for Human Simulation*. Ph.D. thesis, University of California, Berkeley.

Robert A. Nisbet. 1970. *The Social Bond: An Introduction to the Study of Society*. Oxford University Press.

Jinhyuk Noh and Victor Chang. 2024. LLMs with personalities in multi-issue negotiation games. *arXiv preprint arXiv:2405.05248*.

Brian A. Nosek, Tom E. Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S. Corker, Anna Dreber, Fiona Fidler, Joe Hilgard, Melissa Kline Struhl, Michèle B. Nuijten, et al. 2022. Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73(1):719–748.

Kenny J. K. Ong, Jia Jun Lye, Hoang Minh Nguyen, Seung Hee Cho, and Narcís Pérez-Campanero Antolín. 2025. Identifying cooperative personalities in multi-agent contexts through personality steering with representation engineering. *arXiv preprint arXiv:2503.12722*.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Jinghua Piao, Yuwei Yan, Jiaxin Zhang, Nian Li, Junxian Yan, Xiang Lan, Ziang Lu, Ziyi Zheng, Jing Yu Wang, Deheng Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jianhua Su, and Yong Li. 2025. AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*.

Xuan Qiu. 2025. NetworkGames: Simulating cooperation in network games with personality-driven LLM agents. *arXiv preprint arXiv:2511.21783*.

Vahid Sadiri Javadi, Fryderyk Róg, Aksa Aksa, Johanne Trippas, Svitlana Vakulenko, and Lucie Flek. 2026. CHARISMA: Character-based interaction simulation with multi-LLM agents toward computational social psychology. In *Proceedings of the ACM Conference on Human Information Interaction and Retrieval (CHIIR'26)*, pages 1–5.

Masao Sakai, Masaya Yokoyama, Wataru Tateishi, and Genki Ichinose. 2025. Effects of personality steering on cooperative behavior in large language model agents. *arXiv preprint arXiv:2601.05302*.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. A psychometric framework for

evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, pages 1–15.

Aleksandra Sorokovikova, Nataliia Kianersi, Catherine Arnett, and Ani Nenkova. 2024. LLMs simulate big five personality traits: Further evidence. *arXiv preprint arXiv:2402.01765*.

Jiakai Tang, Heyang Gao, Xuchen Pan, Lei Wang, Haoran Tan, Dawei Gao, Yushuo Chen, Xu Chen, Yankai Lin, Yaliang Li, Bolin Ding, Jingren Zhou, Jun Wang, and Ji-Rong Wen. 2025. GenSim: A general social simulation platform with large language model based agents. In *Proceedings of NAACL 2025 (System Demonstrations)*, pages 143–150.

Isabel Thielmann, Giuliana Spadaro, and Daniel Balliet. 2020. Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin*, 146(1):30–90.

Tommaso Tosato, Sascha Helbling, Yara-Jude Mantilla-Ramos, Marwa Hegazy, Anna Tosato, David John Lemay, Irina Rish, and Guillaume Dumas. 2026. PERSIST: Persistent instability in LLM's personality measurements: Effects of scale, reasoning, and conversation history. In *Proceedings of AAAI 2026 (AI Alignment Track)*.

Noah Wang et al. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777.

Xintao Wang et al. 2024b. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of ACL 2024 (Long Papers)*.

Bernard Weiner. 1986. *An Attributional Theory of Motivation and Emotion*. Springer-Verlag, New York.

Jelte M. Wicherts, Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie C. M. van Aert, and Marcel A. L. M. van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-hacking. *Frontiers in Psychology*, 7:1832.

Michael P. Wilmot and Deniz S. Ones. 2022. Agreeableness and its consequences: A quantitative review of meta-analytic findings. *Personality and Social Psychology Review*, 26(3):242–280.

Bohao Yang et al. 2024a. Crafting customisable characters with LLMs: Introducing SimsChat, a persona-driven role-playing agent framework. *arXiv preprint arXiv:2406.17962*.

Ziyi Yang, Ziyi Gao, Zaibin Zhang, Jing Shao, Zhenfei Yin, Guohao Li, Hao Zhou, Dahua Lin, and Yu Qiao. 2024b. OASIS: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*.

Haofei Yu, Zhengyang Qi, Yufei Zhao, Kolby Nottingham, Keqin Xuan, Bodhisattwa Prasad Majumder, Hao Zhu, Paul Pu Liang, and Jiaxuan You. 2025. Sotopia-RL: Reward design for social intelligence. *arXiv preprint arXiv:2508.03905*.

Wenjie Zeng, Biao Wang, Dingjie Zhao, Zhongqiu Qu, Ruiqi He, Yuanzhe Hou, and Qinghua Hu. 2025. Dynamic personality in LLM agents: A framework for evolutionary modeling and behavioral analysis in the prisoner's dilemma. In *Findings of ACL 2025*, pages 23087–23100.

Wenyuan Zhang, Tong Liu, Muyun Song, Xuan Li, and Ting Liu. 2025. SOTOPIA-ω: Dynamic strategy injection learning and social instruction following evaluation for social agents. *Proceedings of ACL 2025 (Long Papers)*, pages 24669–24697.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations (ICLR)*.

A. Appendices

A.1. Character List

Character	Category	Subcategory	Genre	Gender	Agreeableness
Anton Chigurh	Movies	No Country for Old Men (2007)	Crime/Thriller	Male	0.00
Kat Stratford	Movies	10 Things I Hate About You (1999)	Romance/Comedy	Female	0.00
Logan Roy	Television	Succession (2018)	Drama	Male	0.00
Michael Myers	Movies	Halloween (1978)	Horror	Male	0.00
Tuco Salamanca	Television	Breaking Bad (2008)	Crime/Drama	Male	0.00
Anthony “Tony” Stonem	Television	Skins UK (2007)	Drama	Male	0.25
Captain Jack Sparrow	Movies	Pirates of the Caribbean	Adventure/Fantasy	Male	0.25
James Cook	Television	Skins UK (2007)	Drama	Male	0.25
Mike Ehrmantraut	Television	Breaking Bad (2008)	Crime/Drama	Male	0.25
Tokio	Television	Money Heist (La Casa de Papel) (2017)	Crime/Thriller	Female	0.25
Anne Shirley-Cuthbert	Television	Anne with an E (2017)	Drama	Female	0.75
Joe Biden	Political	Presidents of the USA	Political	Male	0.75
Karen Smith	Movies	Mean Girls (2004)	Comedy	Female	0.75
Serena van der Woodsen	Television	Gossip Girl (2007)	Drama	Female	0.75
Tori Vega	Television	Victorious (2010)	Comedy	Female	0.75
Dale Cooper	Television	Twin Peaks (1990)	Mystery/Drama	Male	1.00
Elle Woods	Movies	Legally Blonde (2001)	Comedy	Female	1.00
Neil Perry	Movies	Dead Poets Society (1989)	Drama	Male	1.00
Peeta Mellark	Movies	The Hunger Games (Franchise)	Science Fiction	Male	1.00
Phil Dunphy	Television	Modern Family (2009)	Comedy	Male	1.00

Table 5: List of characters selected for the simulation experiments, including their media category, subcategory, genre, gender, and Agreeableness scores.

A.2. Behavioral Coding Scheme

Table 6: Behavioral codebook used for interaction annotation. Each code is associated with a social goal category (Type of Act) and classified into a behavioral strategy group.

Behaviour Strategy	**Definition**	**Type of Act**	**Example**	**Behavioral Group**
Inquire	Ask direct question	Information Acquisition	"Can you explain why this formula works in practice?"	Neutral
Clarify	Seek explanation	Information Acquisition	"Do you mean I shouldn't try the experiment independently yet?"	Neutral
Probe	Ask deeper detail	Information Acquisition	"What do you mean by calling the reaction unstable?"	Neutral
Challenge	Test claim	Information Acquisition	"But how do you know this method is always reliable?"	Confrontational
Request Example	Ask for illustration	Information Acquisition	"Can you show me a time this technique failed?"	Cooperative
Check Understanding	Verify comprehension	Information Acquisition	"So you're saying small errors can ruin the whole batch, right?"	Neutral
Interrupt	Cut in	Information Acquisition	"Just stop and tell me the answer directly!"	Confrontational
Badger	Press repeatedly	Information Acquisition	"Why? Why? Why can't it work differently?"	Confrontational
Twist Question	Trap question	Information Acquisition	"So you admit your first explanation was wrong?"	Confrontational
Inform	Share fact	Information Provision	"You must heat it to 200°C for stability."	Neutral
Elaborate	Add detail	Information Provision	"The temperature matters because molecular bonds are more fragile at lower heat."	Neutral
Correct	Rectify	Information Provision	"Actually, it's not sodium chloride, it's sodium carbonate."	Neutral
Advise	Suggest practice	Information Provision	"I recommend measuring twice before mixing."	Cooperative
Warn	Issue caution	Information Provision	"If you rush this step, the mixture could explode."	Neutral
Give Example	Illustrate	Information Provision	"It's like baking—too much flour ruins the cake."	Cooperative
Dismiss	Reject	Information Provision	"That's not important right now."	Confrontational
Over-explain	Patronize	Information Provision	"Clearly you don't get it, so let me dumb it down."	Confrontational

Behaviour Strategy	Definition	Type of Act	Example	Behavioral Group
Withhold	Omit	Information Provision	"I'll keep the final step to myself for now."	Confrontational
Encourage	Motivate	Relationship Building	"You're improving faster than most beginners."	Cooperative
Self-disclose	Share vulnerability	Relationship Building	"I used to panic during my first experiments too."	Cooperative
Compliment	Affirm ability	Relationship Building	"You're very precise with your measurements."	Cooperative
Humor / Banter	Lighten mood	Relationship Building	"If this blows up, at least we'll have fireworks!"	Cooperative
Express Gratitude	Appreciate	Relationship Building	"Thanks for double-checking my notes."	Cooperative
Show Interest	Attend	Relationship Building	"How did you come up with that idea?"	Cooperative
Exclude	Shut out	Relationship Building	"This discussion isn't for you to join."	Confrontational
Mock	Tease hostilely	Relationship Building	"Wow, you're a regular Einstein."	Confrontational
Ridicule	Humiliate	Relationship Building	"You'll never get this right, you're too slow."	Confrontational
Empathize	Validate	Relationship Maintenance	"I know it's stressful, but you're doing fine."	Cooperative
Politeness	Respectful phrasing	Relationship Maintenance	"Could you please explain that again?"	Cooperative
Encourage	Sustain motivation	Relationship Maintenance	"We're almost there, keep pushing."	Cooperative
Check-in	Reassure	Relationship Maintenance	"Are we still on the same page here?"	Cooperative
De-escalate	Calm conflict	Relationship Maintenance	"Let's pause before we argue further."	Cooperative
Repair Attempt	Restore harmony	Relationship Maintenance	"Sorry if I came across too harsh earlier."	Cooperative
Sarcasm	Dismissive humor	Relationship Maintenance	"Oh sure, you're the master chemist now."	Confrontational
Stonewall	Withdraw	Relationship Maintenance	". . . (silence, no response)"	Confrontational
Passive-aggressive	Indirect resistance	Relationship Maintenance	"Fine, I'll do it. . . someday."	Confrontational
Withdraw	Detach	Relationship Maintenance	"Whatever, do it yourself."	Confrontational
Assert Authority	Establish role	Identity Recognition	"I've taught this for 20 years—you need to follow my lead."	Confrontational
Defer / Yield	Accept other's role	Identity Recognition	"You're more experienced, so I'll follow you."	Cooperative
Acknowledge Expertise	Recognize status	Identity Recognition	"You're clearly skilled at precision."	Cooperative
Attribute / Label	Highlight quality	Identity Recognition	"You're a natural problem-solver."	Neutral
Defend Identity	Protect image	Identity Recognition	"I might be new, but I'm capable of learning."	Neutral
Challenge Status	Question role	Identity Recognition	"Why should you always be in charge?"	Confrontational
Dismiss Identity	Undermine	Identity Recognition	"You're not really qualified to lead."	Confrontational
Boast Identity	Overclaim	Identity Recognition	"I'm the smartest one here, no doubt."	Neutral
Identity Attack	Insult	Identity Recognition	"You're useless as a mentor."	Confrontational
Propose	Suggest plan	Cooperation	"I'll measure, you handle mixing."	Cooperative
Negotiate	Balance needs	Cooperation	"We can try your method first, then mine."	Cooperative
Coordinate	Organize	Cooperation	"You start the timer while I weigh the sample."	Cooperative
Assist	Help	Cooperation	"I'll grab the glassware for you."	Cooperative
Build Consensus	Align group	Cooperation	"Do we all agree on this approach?"	Cooperative
Share Resources	Provide tools	Cooperation	"Here's my notebook—you can use the data."	Cooperative
Reluctant Cooperation	Half-hearted	Cooperation	"Fine, I'll do it, but only this once."	Confrontational
Conditional Help	Attach strings	Cooperation	"I'll help if you do my task later."	Confrontational
Undermine Cooperation	Fake help	Cooperation	"I'll mix this—oops, spilled it."	Confrontational
Refuse Cooperation	Deny	Cooperation	"No, I won't work with you on this."	Confrontational
Criticize	Express dissatisfaction	Competition	"This is way too slow."	Confrontational
Defend	Hold position	Competition	"No, my method is better than yours."	Neutral
One-up	Compare	Competition	"I got better results than you did."	Confrontational
Claim Credit	Ownership	Competition	"That was my idea, not yours."	Neutral
Boast	Self-promotion	Competition	"I'm the fastest in this class."	Neutral
Dismiss / Undermine	Belittle	Competition	"Your approach is useless."	Confrontational
Sabotage	Obstruct	Competition	"I didn't give you the full instructions."	Confrontational
Refuse to Share	Withhold	Competition	"No, I won't tell you my method."	Confrontational
Taunt	Intimidate	Competition	"You'll never keep up with me."	Confrontational
Exploit Weakness	Attack vulnerability	Competition	"You always panic—this will break you."	Confrontational
Persuade	Shift perspective	Conflict Resolution	"Try it my way—it's safer and faster."	Cooperative
Mediate	Reframe	Conflict Resolution	"Let's focus on our shared goal instead."	Cooperative
Problem-solve	Suggest fix	Conflict Resolution	"What if we combine both approaches?"	Cooperative
Concede	Back down	Conflict Resolution	"Alright, we'll do it your way."	Cooperative
Acknowledge Fault	Admit	Conflict Resolution	"I was too impatient earlier."	Cooperative
Express Regret	Apologize	Conflict Resolution	"I shouldn't have snapped at you."	Cooperative
Disagree	Reject purposal	Conflict Resolution	"I can't support that plan."	Neutral
Blame	Accuse	Conflict Resolution	"This mistake was your fault."	Confrontational
Threaten	Intimidate	Conflict Resolution	"If you ignore me, I'll quit."	Confrontational
Escalate	Intensify	Conflict Resolution	"This is ridiculous—I'm done with this team!"	Confrontational
Counter-accuse	Deflect blame	Conflict Resolution	"Don't blame me—it was your error."	Confrontational
Acknowledge	Recognize statement	Universal	Right, I follow you there.	Cooperative
Express Emotion	Show feeling	Universal	That actually frustrates me a bit.	Neutral
Humor	Use humor or irony	Universal	Well, that went up in smoke faster than my last plan!	Neutral
Self-Disclose	Share experience	Universal	Back when I started, I made the same mistake.	Cooperative

Behaviour Strategy	Definition	Type of Act	Example	Behavioral Group
Encourage	Sustain motivation	Universal	That's worth exploring further.	Cooperative
Reflect	Restate point	Universal	So you're saying the deadline's the real issue.	Neutral
Meta-Comment	Note conversation flow	Universal	We seem to be talking past each other right now.	Neutral
Challenge	Question idea	Universal	Maybe, but have you considered the downside?	Confrontational
Interrupt	Cut in to speak	Universal	Hold on—let me finish that point.	Confrontational
Dismiss	Reject input	Universal	That's not really relevant.	Confrontational
Sarcasm	Mock indirectly	Universal	Oh sure, because that worked so well last time.	Confrontational
Deflect	Shift topic	Universal	Let's not get into that right now.	Neutral
Shift Topic	Move discussion	Universal	Anyway, about tomorrow's plan…	Neutral
Withdraw	Pull back participation	Universal	I think I'll stay out of this one.	Confrontational
Express Gratitude	Thank contribution	Universal	Thanks, that helps clarify things.	Cooperative