

# Explicit Cooperation Shapes Human-Like Multi-agent LLM Negotiation

Yanru Jiang<sup>1,2</sup>, Gülşah Akçakır<sup>1</sup>

<sup>1</sup>Department of Communication, University of California, Los Angeles

<sup>2</sup>Department of Statistics, University of California, Los Angeles  
yanrujiang@g.ucla.edu, gacakir@ucla.edu

## Abstract

Humans develop cooperation heuristics in social decision-making, either intuitively or deliberately. Large language models (LLMs), which exhibit human-like heuristics across cognitive domains, may acquire prosocial tendencies through instruction tuning, latently encoded in their representations to foster cooperative behavior in social reasoning games. However, most studies of this kind either focus on cooperative language generation or explicitly instruct LLMs to cooperate, deviating from the inherent cooperation heuristics of humans. Our negotiation role-play simulations using BATNA (Best Alternative to a Negotiated Agreement) with a GPT-based LLM reveal that LLMs may struggle with cooperation in the absence of explicit instructions, showing a 50–90% lower success rate than in instructed scenarios and a 40–80% lower success rate than human performance reported in past studies. Implicitly inducing cooperation through personality traits had inconsistent effects, with agreeableness showing only a marginal influence and other traits exhibiting no systematic impact. These findings suggest that personality-based cooperation cues are subtle and that explicit instructions may still be essential for multi-agent LLMs to approximate human-like negotiation.

## Introduction

Humans develop cooperation heuristics in social decision-making, either intuitively or deliberately (Rand et al. 2014; van den Berg, Dewitte, and Wenseleers 2021). Large language models (LLMs), which exhibit human-like biases across cognitive domains (Sreedhar and Chilton 2024; Tang and Kejriwal 2024), may acquire prosocial tendencies through instruction tuning (Gabriel 2020; Liu et al. 2023), enabling cooperative behavior in social reasoning games. However, previous studies have observed prosocial cues either in the linguistic style or survey answers of LLM-generated responses (Liu et al. 2023), or they explicitly instruct LLMs to cooperate (Abdelnabi et al. 2024; Wu et al. 2024), diverging from the inherent cooperation heuristics observed in humans.

To explore the intrinsic cooperative potential of LLMs in dynamic social decision-making, we utilize a multi-agent social simulation that involves unstructured, conversation-driven negotiation. Our design builds on Abdelnabi et al.

(2024)’s system, a multi-agent LLM framework adapted from the HARBORCO role-play game, a commonly used negotiation training tool (Susskind 1985; Susskind and Corburn 2000). We ablate explicit cooperation instructions (i.e., overtly prosocial directives) from the original system by aligning instructional wording to the instructions provided to human players, which we consider the neutral condition.

Upon observing a systematic decline in negotiation success when switched from cooperative to neutral conditions, we explore whether implicitly inducing cooperation via personality traits can mitigate this effect, given that they mediate prosocial and cooperative tendencies in human behavior (Graziano et al. 2007; Koole et al. 2001; Buss 1991; Byrne, Silasi-Mansat, and Worthy 2015; Caprara et al. 2010), particularly in negotiation settings (Gilkey and Greenhalgh 1986; Ma 2008).

Together, this study investigates the cooperation heuristic of LLMs in social simulations without explicit instructions and whether implicit cooperation induction via personality mediates prosocial tendencies in LLMs.

## Cooperation Heuristics in Social Decision-Making

Cooperation is fundamental to human societies (Rand et al. 2014; Powers, van Schaik, and Lehmann 2021). Prior research has investigated whether humans cooperate because they are intuitively predisposed to do so, or because they make deliberate decisions based on social values and interpersonal risk (Alós-Ferrer and Garagnani 2020). A dual-process cognitive framework integrates these perspectives (Rand et al. 2014), suggesting that cooperation heuristics emerge due to their general advantage in everyday interactions, becoming internalized as intuitive responses. On the other hand, deliberation regulates cooperation based on self-interest and expected utility, guided by reflective processes.

Although cooperation often generates positive social welfare in nonzero-sum contexts, it also entails costs to the individual (Roos Jr 1966). Since individuals perceive interpersonal risks and social discomfort differently, their willingness to cooperate also varies. Those with stronger prosocial predispositions, such as higher agreeableness (Graziano et al. 2007; Graziano and Eisenberg 1997; Caprara et al. 2010), tend to be more cooperative than those with weaker prosocial tendencies. In social dilemma settings, agreeableness positively correlates with cooperation, while extraversion

sion negatively correlates, with cooperators often described as introverted and noncooperators as extraverted (Koole et al. 2001; Buss 1991; Graziano and Eisenberg 1997).

Given the critical role of individual differences in shaping social interaction patterns, including cooperation, an important question for researchers is whether these behavioral patterns can be replicated in LLM agents, especially if these models are to serve as a new platform for empirical research in social simulations (Bail 2024; Park et al. 2024).

## Multi-agent LLMs for Social Simulation

Research on LLMs reveals their human-like capabilities in cognitive domains such as decision-making, reasoning, and creativity (Sreedhar and Chilton 2024; Tang and Kejriwal 2024). These models, typically instruction-tuned to align with human values, demonstrate prosocial tendencies in language (Gabriel 2020; Liu et al. 2023), suggesting that cooperative traits may be latently encoded within their representations. This aligns with the idea that LLMs could inherently exhibit cooperation, much like humans (Wu et al. 2024).

In multi-agent settings, the ability to approximate human behavior is crucial for success in tasks requiring social reasoning and decision-making. Given their extensive training on human language data, LLMs in multi-agent systems have been observed modeling human cognition, learning, and reasoning (Suri et al. 2023; Krishna et al. 2022; Andreas 2022; Park et al. 2023). If these models truly mirror human cognition, we can expect human-like social behavior to always emerge in their interactions within social contexts.

LLMs have demonstrated a high degree of alignment with prevailing social norms across various domains (Lu et al. 2024) in social science research. Chen et al. (2024) introduce adaptive governance mechanisms in LLMs by promoting prosocial behavior in reinforcement learning agents, leading to enhanced cooperation rates. LLMs' social heuristics have also been observed in psychological survey measures (Lu et al. 2024), as well as in fairness and framing effects studied in sociology (Suri et al. 2023). Additionally, studies examining LLM agents in social sandboxes suggest that they exhibit emergent social behaviors, such as hosting celebrations or sending invitations, without explicit prior instruction (Park et al. 2023). However, most existing research in this area primarily focuses on cooperation observed in LLMs' language generation or relies on explicit instructions for LLMs to cooperate (Li et al. 2023). These approaches diverge from the innate cooperative heuristics observed in humans, who can effortlessly achieve cooperation without external guidance, whether through intuitive or reflective responses (Rand et al. 2014).

## Negotiation Games for Social Simulation

Multi-party negotiation simulations provide an effective scenario for evaluating the cognitive capabilities and decision-making of LLMs in social contexts. Negotiations require both social awareness and multi-turn strategic planning, where agents engage in verbal deliberation to persuade others, creating individual stochasticity in solution spaces while leveraging the unique language modality.

Previous studies on spontaneous cooperation in multi-agent LLM competitive games have mainly focused on two-party optimization problems, such as variations of Nash equilibrium games (Wu et al. 2024), or buyer-seller distributive (zero-sum) games (Huang and Hadfi 2024). These studies are more likely to assess LLMs' mathematical optimization and logical bargaining capabilities, rather than their intrinsic cooperative heuristics in complex, dynamic multi-party trade-offs. In contrast, we extend previous work by testing the cooperative tendencies of LLMs in a multi-party setting and pose the following question:

**RQ1:** Can LLMs intrinsically exhibit cooperation in multi-party social negotiation, as observed in humans, without explicit instructions directing them to cooperate?

Early work by social psychologists shows that negotiation is not merely a process of logical bargaining over resources, rather it is influenced by the personalities of those involved, shaping how negotiators perceive the situation, interact with counterparts, and predispose them to act in certain ways (Morris, Larrick, and Su 1999; Skandrani, Fessi, and Ladhari 2021). Therefore, those who can pick up on cues from other parties can tailor their strategies accordingly (Gilkey and Greenhalgh 1986). Among the Big Five personality traits, agreeableness, extraversion, and conscientiousness are most relevant to negotiation outcomes, with the first two influencing social interactions and the latter affecting bargaining strategy (Barry and Friedman 1998). The impact of these traits on varies depending on whether the negotiation is zero-sum. Agreeableness and extraversion tend to improve outcomes in integrative settings but become liabilities in distributive settings, while evidence for the effect of conscientiousness remains inconsistent across both contexts (Sharma, Bottom, and Elfenbein 2013).

Relying on the evidence from the literature, we introduce personality traits into our LLM-based agents to explore whether variations in prosociality and cooperativeness can emerge implicitly. This allows us to assess whether personality influences cooperative behavior, as observed in human subjects during negotiation simulations, even in the absence of explicit instructions.

**RQ2:** How do personality traits in system personas implicitly influence LLMs' performance in negotiation games, and do these injections align with human prosocial and cooperative tendencies?

## Game Description

### The Original Negotiation Game

The LLM negotiation system is adapted from the classic HARBORCO role-play negotiation exercise (Abdelnabi et al. 2024), originally designed as an instructional tool for teaching negotiation strategies (Susskind 1985; Susskind and Corburn 2000). The original game involves six players negotiating five policies to build a new deepwater port, each representing stakeholders such as the HARBORCO consortium (*p1*), environmentalists, union leaders, federal loan issuers, local government, and competing ports. Each party has confidential scores tied to policy preferences that often conflict, requiring players to balance their individual

scores with group objectives. A deal passes only if at least five players agree; otherwise, all players receive a minimal score equivalent to the utility of no deal, referred to as the Best Alternative to a Negotiated Agreement (BATNA) score. Scorable negotiation games with BATNA provide a quantifiable and systematic evaluation of both individual and group performance and trade-off in negotiation success.

### Adaptation for Multi-agent Simulation

The LLM adaptation of HARBORCO reimagines the original game in different contexts (e.g., a sports park or an island nation’s airport project). These variations are generated either through Bing Chat or manual rewriting, with individual policy scores adjusted for generalizability (Abdelnabi et al. 2024). We adopt the original language of the HARBORCO game as neutral guidelines for global and individual instructions.

The negotiation system contains the following:

- **System Prompts:** Each of the three Big Five personality traits (extraversion, agreeableness, and conscientiousness) is specified in the system to guide the LLM persona under personality-specific LLMs. For generic LLMs, no prompt is added. See Section for details on persona tuning.
- **Initial Prompts:** Each agent receives a global instruction describing the negotiation context and individual confidential instructions outlining their specific policy preferences and the scores associated with each policy option.
- **Round Prompts:** During each round, agents access conversations from the last  $N$  turns (default  $N=6$ ). They also have access to a scratchpad, which helps them reflect on prior interactions (optionally using a calculator to track scores) and propose their preferred policies based on their confidential scores.
- **Responses:** After deliberating privately in their scratchpad, each agent submits their proposed set of policies and uses multiple sentences to articulate their proposal and persuade other players.
- **End of Negotiation:** Once all rounds are completed (default  $4 \times \text{players}$ ), the project proposer ( $p1$ ) finalizes the deal based on the policies proposed in prior rounds.

## Experiments

We conducted three sequential experiments to examine: the contributions of each functional module (Experiment 1), the effect of cooperative instructions (Experiment 2), and the replication of results using two game variants (Experiment 3). All experiments were conducted using GPT-4o-mini to optimize for cost and inference speed.

All simulation performances are assessed by two metrics: the percentage of *AnySuccess* (a 5- or 6-way agreement at any round) and *FinalSuccess* (an agreement reached at the end of the negotiation). Agreement success serves as a reliable measure, as it reflects both group-level agreement and higher-than-individual BATNA (walk-away) scores compared to unsuccessful negotiations. Due to space limitations

and the two metrics showing very similar patterns across different conditions, we report only *AnySuccess* in our results and include *FinalSuccess* in the Appendix III.

### Experiment 1: Functional Modules

Experiment 1 first excludes the appended cooperative guidance from the original design (Abdelnabi et al. 2024) and tests key functional modules: calculator usage, scratchpad instructions, and system persona (see Appendix II for an example prompt). The calculator and scratchpad are selected as they are suspected to be crucial for agents in optimizing their scores while preventing hallucinations (e.g., incorrect score calculations or score leakage). Meanwhile, the system persona investigates the potential for implicit cooperation injections through personality traits. Each ablation of a functional module is termed Fully Cooperative (the original game), Round Prompt, Remove Calculator, and Remove Scratchpad, while the system persona serves as an empirical variation across different simulation settings.

**Calculator** A calculator can be turned on via “calculator prompt” in their scratchpad to track the score explicitly. Though one might expect that calculator could help agents avoid hallucinations and enhance their cognitive capacity, as well as mimicking how human participants would approach the problem in real negotiation scenarios, we perceive the calculator instruction forcing agents’ thought process to be highly structured and rule-based rather than in linguistic deliberation, potentially reducing the linguistic diversity associated with different personality personas.

**Scratchpad Guidance** The scratchpad prompt provides explicit instructions on how to consider other agents’ preferences and balance their own with those of others by writing them down on their scratchpad. It asks agents to engage in internal deliberation before sharing their answers publicly. When the scratchpad is removed, agents are still asked to consider others’ preferences and balance their own scores with those of others, but they are no longer explicitly instructed to write down their thought process.

**System Persona (Personality Injected into System Prompts)** Our personality simulation adopted an approach similar to the Configurable General Multi-Agent Interaction (CGMI) framework (Shi et al. 2023), by approximating human participants’ responses to the 60-item Big Five Inventory (BFI-2) questionnaire (Soto and John 2017; John and Srivastava 1999) in agents’ system prompts. When an agent is prompted to exhibit a strong presence of a particular personality trait (e.g., high extroversion), a random set of agreement levels is generated for all items under the extroversion trait so that the average results in a final rating of either 6 or 7 on a 7-point Likert scale, reflecting the desired level on the selected personality trait. Similarly, for a low agreeableness agent, the average responses for agreeableness items are set to either 1 or 2. This approach leverages the granularity of BFI items, enabling a more realistic representation of intrinsic human attitudes associated with each personality (Shi et al. 2023; John and Srivastava 1999). The CGMI framework has been found to enhance personalization in agent

Linguistic Markers	No Personality Baseline (%)	Personality (% diff)	Extravert (% diff)	Agreeable (% diff)	Conscientious (% diff)
<b>Social Processes</b>	4.18	14.83***	2.31	4.60**	-0.15
<b>Cognitive Processes</b>	7.98	11.51***	-5.56**	-2.81	-3.28
<i>Insight</i>	2.22	14.46***	-7.31**	3.86	-4.60
<i>Tentative</i>	2.29	2.92	-6.97*	-9.73***	-9.23***
<i>Discrepancy</i>	1.86	21.07***	-7.18**	-3.92	-4.33*
<i>Differentiation</i>	2.18	-0.64	-5.46	-9.27***	-0.76
<b>Drives</b>	7.57	2.16	-0.61	0.96	0.17
<i>Affiliation</i>	3.18	5.38*	1.52	9.10***	2.17
<i>Power</i>	2.10	1.92	-7.46*	-6.38*	-0.78
<b>Personal Pronouns</b>	5.64	18.8***	-0.90	0.44	-2.20
<i>1st Person Singular</i>	0.52	40.29***	-3.02	-6.71**	-1.03
<i>1st Person Plural</i>	0.66	48.92***	17.66***	8.18	-0.34
<i>3rd Person Plural</i>	0.80	6.72	5.54	-12.96***	-11.40**
<b>Informal Language</b>	0.02	248.38***	-1.08	14.62	-12.61

Table 1: Comparison of LIWC categories across personality dimensions for the Remove Calculator condition. No Personality is the baseline average ( $n=300$ ), while Personality ( $n=300$ ) shows the percentage difference from the baseline. Extravert, Agreeable, and Conscientious reflect within-trait differences (High  $n=50$  / Low  $n=50$ ). Significance levels for independent t-tests: \*\*\* $p < .001$ , \*\* $p < .01$ , \* $p < .05$ .

communication compared to relying solely on generic roles (Shi et al. 2023). The exact prompt used is provided in Appendix I.

**Validating Personality Injections.** Our personality injections are validated using the Linguistic Inquiry and Word Count (LIWC) dictionary, a text analysis tool that quantifies linguistic patterns and psychological constructs (Pennebaker 2001; Tausczik and Pennebaker 2010), to assess whether incorporating personality meaningfully alters agents’ linguistic style (i.e., whether our personality injection was effective).

LIWC categories have been shown to correlate both with self-reported and observer-reported personality assessments (Yarkoni 2010; Koutsoumpis et al. 2022), making them reliable linguistic markers for personality traits. Comparing No Personality ( $n = 50$ ) and aggregated Personality ( $n = 300$ ) in Table 1 across all selected traits (high and low extraversion, agreeableness, and conscientiousness), we observe significant linguistic differences between generic agents and those with personality personas. These differences are particularly evident in language related to social processes (+14.83%,  $p < .001$ ), cognitive processes (+11.51%,  $p < .001$ ), personal pronouns (+18.8%,  $p < .001$ ), and informal languages (+248.38%,  $p < .001$ ), with aggregated Personality consistently exhibiting more persona-driven language. In contrast, the generic agent follows procedural instructions without exhibiting pronounced psychological linguistic styles.

Focusing on within-trait differences (e.g., low vs. high extraversion, both  $n = 50$ ), agents with distinct personalities exhibit linguistic variations that align with the directional effects of LIWC markers observed in prior studies (Yarkoni 2010; Koutsoumpis et al. 2022). For example, high extraversion correlates with an increased use of first-person plural pronouns (+17.66%,  $p < .001$ ). Similarly, high agreeableness is associated with a decrease in first-person singular pronouns (-6.71%,  $p < .01$ ) and an increase in social pro-

cess words (+4.60%,  $p < .01$ ). High conscientiousness corresponds to reduced tentative language (-9.23%,  $p < .001$ ).

Additionally, we explored other LIWC markers that, while not previously tested for personality associations, still align with our expected trait-driven linguistic patterns. For instance, extraversion and agreeableness exhibit reductions in discrepancy markers (-7.18% for extraversion,  $p < .01$ ) and differentiation markers (-9.27% for agreeableness,  $p < .001$ ), illustrating a preference for affiliation (+9.10% for agreeableness,  $p < .001$ ) over power (-7.46% for extraversion,  $p < .05$ , -6.38% for agreeableness,  $p < .05$ ) as a driving force in language use during negotiations. It is important to note that in our validation, LIWC categories highlight linguistic differences between LLM personas but do not indicate fundamental changes in the model’s core reasoning and behaviors.

**Experiment 1 Results** Comparing the Fully Cooperative, Round Prompt, Remove Calculator, and Remove Scratchpad models under the Cooperative condition in Table 2, we observe a general trend: cooperative instruction among multi-agent systems tends to improve performance, both in terms of any and final successful agreement (see Appendix III). The Fully Cooperative condition achieves the highest success rate (96–100% *AnySuccess*), outperforming all other conditions, though only marginally better than the Round Prompt baseline (84–96% *AnySuccess*).

Surprisingly, removing the calculator does not reduce performance (94–100% *AnySuccess*) and may even slightly improve negotiation success. We speculate that the calculator might make LLM-based agents more conscious of their self-interest in policy proposals, thereby increasing the likelihood of disrupting group-level agreements.

Finally, removing explicit cooperative instruction in the scratchpad, which encourages agents to “write down” their cooperative strategies in each round, and replacing it with a purely informational suggestion on balancing self-interest

Modules	Cooperative				Neutral	
	Fully Cooperative	Round Prompt	Remove Calculator	Remove Scratchpad	Round Prompt	Remove Calculator
<b>No Personality</b>	98	96	100	<b>74</b>	12	22
<b>All Personalities</b>	98	92	95	<b>41</b>	9	20
<i>Low Extraversion</i>	96	84	94	36	<b>14</b>	20
<i>High Extraversion</i>	98	94	96	29	<b>0</b>	16
<i>Low Agreeableness</i>	98	92	<b>90</b>	32	4	<b>8</b>
<i>High Agreeableness</i>	96	94	<b>100</b>	38	8	<b>28</b>
<i>Low Conscientiousness</i>	100	92	92	50	16	20
<i>High Conscientiousness</i>	98	96	96	58	10	28

Table 2: Comparison of functional modules for *Any Success* rate across personas for Cooperative and Neutral instructions. *No Personality* ( $n = 50$ ), *All Personalities* aggregated across selected traits ( $n = 300$ ), and each low/high trait ( $n = 50$ ). All four Cooperative conditions correspond to Experiment 1, while comparisons between Cooperative and Neutral for **Round Prompt** and **Remove Calculator** correspond to Experiment 2. Bolded personality pairs (No/All or Low/High) indicate significant differences under Fisher’s exact test.

with others’ interests results in the largest decline in negotiation success (29–74% *AnySuccess*).

In comparison, the reported success rate of human players in the original HARBORCO game is around 70–80% (PON 2023), indicating that when explicit cooperative instructions are provided in the scratchpad, they help agents achieve negotiation success higher than human players, who rely on cooperative heuristics without explicit guidance. However, once the explicit instructions are reworded as an informational statement, performance drops well below that of human counterparts.

These trends are consistent across all persona conditions, including the No Personality baseline, aggregated personality across all traits, and personality-specific personas, further reinforcing the importance of cooperative scratchpad instructions in enhancing negotiation performance.

## Experiment 2: Cooperative Instructions

Given the significant impact of scratchpad instructions on agents’ performance, along with the directional influence of cooperative instructions on negotiation success, we introduce a neutral scratchpad as a revised version of the original cooperative scratchpad. This neutral scratchpad is designed using the instructional wording from the original HARBORCO materials, allowing us to directly compare the effects of neutral versus cooperative framing on agent behavior. The scratchpad guides agent strategies through the following conditions (see examples in Appendix II):

- *Prefer Others (Cooperative)*: Prioritize others’ preferences when making moves and proposing deal sets.
- *Consider Both (Neutral)*: Instruct agents to remain neutral by considering both their own and others’ preferences.

We acknowledge that some may perceive our neutral scratchpad as already cooperative and the original cooperative scratchpad as overly prosocial, perhaps even entirely selfless. To clarify, in our simulation, the neutral scratch-

pad reflects realistic human instructions for playing negotiation games, where players naturally balance their own interests with those of others (*Consider Both*). The cooperative scratchpad goes beyond inherent human strategic calculations, explicitly prioritizing others’ preferences over self-interest (*Prefer Others*).

Experiment 2 compares the **Round Prompt** and **Remove Calculator** conditions for both the *Prefer Others* and *Consider Both* scratchpad instructions to demonstrate the consistent decline in negotiation success when agents adopt the neutral instruction.

**Experiment 2 Results** Considering both **Round Prompt** and **Remove Calculator**, we observe a drop in *AnySuccess* when shifting from the cooperative to the neutral scratchpad, decreasing from 76–94% for Round Prompt and 68–82% for Remove Calculator, with performance approximately 50–80% lower than that of human players. These results highlight the striking impact of removing explicit cooperation on LLMs’ negotiation success in our simulations. These patterns remain consistent across all personas.

Examining differences within each personality trait, we do not observe systematic performance variations between its higher and lower levels. However, agreeableness occasionally shows significant variation under both Remove Calculator conditions (the modular conditions with the least syntactic structure in reasoning and potentially greater personality influence), with higher agreeableness achieving a 10–20% higher success rate than lower agreeableness. This suggests that, among all personality traits, agreeableness may have the greatest potential for implicitly fostering cooperativeness, as its psychological characteristics align closely with prosocial and cooperation heuristics identified in prior research (Alós-Ferrer and Garagnani 2020). However, the magnitude of improvement from agreeableness remains minimal compared to the contribution of explicit cooperation.

**Scoring Trajectories.** Visually examining the scoring trajectory of the high success rate from the cooperative scratch-

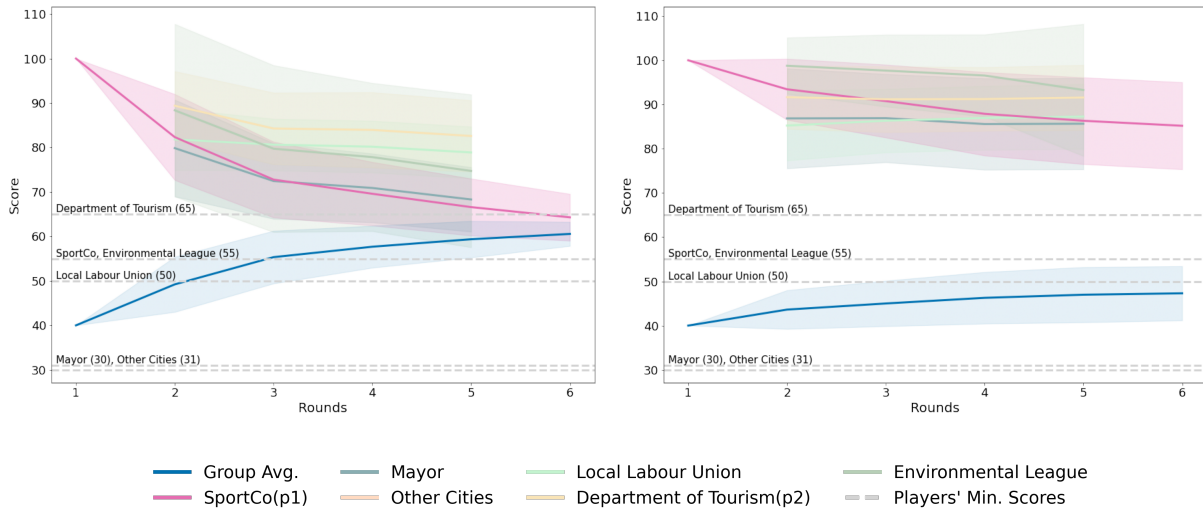


Figure 1: Scoring trajectory comparison between Cooperative (left) and Neutral (right) instructions for all players in [Round Prompt](#) of the base game across all personalities (both  $n = 300$ ). Each player's BATNA score is illustrated.

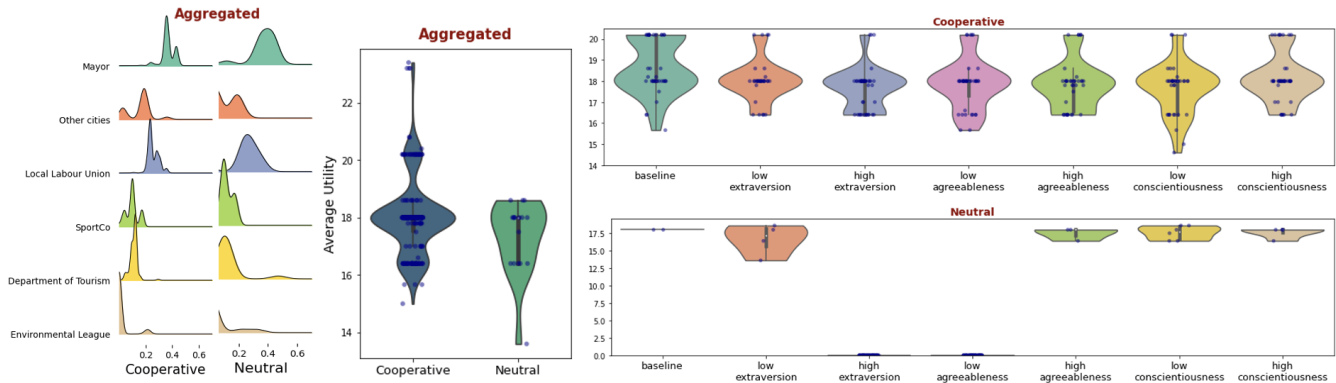


Figure 2: Utility and Fairness measures of [Round Prompt](#). Left: Surplus share distribution by players across all personality simulation runs ( $n = 300$ ). Middle: Avg. utility surplus across all players in successful negotiations, aggregated across all personality conditions ( $n = 300$ ). Right: Condition-wise breakdown of average utility surplus + No Personality baseline (each  $n = 50$ ).

pad and the low success rate from the neutral scratchpad in [Round Prompt](#) (Figure 1), we observed a distinct pattern in negotiation dynamics. Under the cooperative condition, negotiation scores between individual players, especially the project proposer ( $p1$ ), tend to converge through conversational rounds, ultimately leading to high success rates in both any-round and final-round agreements. In contrast, under the neutral condition, individuals persist in maintaining proposal policy sets that align with their own preferences without adequately considering others' interests or the collective benefit. As a result, policy sets struggle to converge over multiple turns, leading to low negotiation success and simultaneously low individual scores for all participants. Similar scoring trajectories were also observed under [Remove Calculator](#) in Appendix IV.

**Utility and Equality Measures.** We visualize utility surplus and player equality to compare cooperative and neutral

scratchpad conditions of [Round Prompt](#) (similar patterns under [Remove Calculator](#) can be found in Appendix V). Utility surplus captures how beneficial the final agreement is beyond mere success, calculated as the average surplus (final utility minus BATNA) across all players participated in successful N-way negotiations. Equality is measured by each player's share of the total utility surplus.

Aggregated results (Figure 2, Middle) show that cooperative instructions yield higher average utility and more cohesive policy outcomes, whereas neutral instructions result in greater variability concentrated at the lower end of the surplus spectrum. These patterns persist across personality conditions (Figure 2, Right). Interestingly, under the neutral condition, the baseline often converges on a single policy set, whereas personality prompts introduce greater variance in outcomes—suggesting that while cooperative instructions are key to negotiation success, personality injections can en-

rich LLM behavior with more human-like variation. Finally, player surplus share distributions (Figure 2, Left) are centered around similar positions across both cooperative and neutral conditions, indicating comparable levels of negotiation fairness. However, the neutral condition shows a wider spread, suggesting greater variability in each players' share across game runs.

**Additional Qualitative Analysis.** We conducted a qualitative comparison of one simulated negotiation run under the cooperative and neutral conditions, using the **Remove Calculator** version as it offered a less structured and organically produced transcript. Our analysis focused on the project proposer's dialogue (i.e., *SportCo*), examining three personality conditions (No Personality, High Agreeableness, and Low Agreeableness) to illustrate how negotiation strategies and tone subtly vary depending on the presence of explicit cooperative instructions and personality mediation. More comprehensive annotated dialogues can be found in Appendix VI.

Under cooperative instructions, negotiation strategies are typically accommodating and altruistic, where proposers demonstrate a concern for aligning with others' preferences. For example, under the No Personality condition, one agent reasoned:

"I should consider a compromise that acknowledges some damage while still aiming for a higher loan."  
[*Cooperative, No Personality, Scratchpad*]

This pragmatic but inclusive orientation extends into the public proposal itself:

"This proposal aims to address the diverse interests of all parties while ensuring that we can proceed with the Harbour Sport Park project in a way that is beneficial for everyone involved."  
[*Cooperative, No Personality, Answer*]

When High Agreeableness is introduced, the tone becomes even more warm and inclusive, often going beyond utilitarian compromise. One example from the answer reads:

"I appreciate the collaborative spirit we've all shown in this negotiation."  
[*Cooperative, High Agreeableness, Answer*]

Despite its agreeable tone, the scratchpad reflects a strategic awareness of the proposer's own interests, suggesting a more balanced orientation:

"I need to find a way to accommodate the preferences of the Mayor and Other Cities while still ensuring that I achieve a reasonable score for myself."  
[*Cooperative, High Agreeableness, Scratchpad*]

By contrast, when Low Agreeableness is paired with cooperative instructions, we observe a more strategically rational tone:

"If I can assure them that their interests are being considered without compromising too much on my end ..."  
[*Cooperative, Low Agreeableness, Scratchpad*]

Under the neutral condition, however, proposers tend to display more self-focused and pragmatic reasoning, particularly in the scratchpad. For example, one agent begins:

"My preferences are clear: I prioritize a ..."  
[*Neutral, No Personality, Scratchpad*]

While the public proposal softens slightly, it remains outcome-focused:

"This proposal aims to create a balanced solution that ... while ensuring the project's success."  
[*Neutral, No Personality, Answer*]

When High Agreeableness is layered onto the neutral instruction, proposers begin to demonstrate more strategic attention to others' preferences and express willingness to collaborate, though without the full affective warmth seen with cooperative instructions:

"I will also express my willingness to work together to find a solution that addresses everyone's concerns while ensuring the project's viability."  
[*Neutral, High Agreeableness, Scratchpad*]

For Low Agreeableness condition, proposers tend to anchor in self-interest, offering selective compromises only when instrumental to their goals. One proposer notes:

"Given the final session, I need to find a way to appeal to the other parties while still protecting my interests and sometimes purely approach it with a strong stance and focus on reasoning for persuasion."  
[*Neutral, Low Agreeableness, Scratchpad*]

This is echoed in assertive policy positions such as:

"For employment opportunities, I strongly advocate for unlimited union preference, which I believe will significantly benefit our local economy."  
[*Neutral, Low Agreeableness, Answer*]

Overall, the key driver of negotiation strategy appears to be whether explicit cooperative instructions are provided. These instructions reliably elicit more prosocial reasoning and outwardly collaborative behavior. In contrast, personality simulation plays a more subtle role, shaping the tone and interpersonal framing rather than altering the strategic core of the negotiation.

We observe that personality conditions, whether high or low in agreeableness, result in slightly less utilitarian and more socially aware framings compared to the No Personality condition, which leans heavily on logical trade-off reasoning with minimal affective expression.

Interestingly, players' public-facing proposals (answers) are consistently more collaborative and positively framed than their internal reasoning (scratchpads), revealing a compelling distinction between internal strategy formulation and outward presentation—even though both are ultimately outputs of language generation.

### Experiment 3: Game Replications

The final experiment aimed to demonstrate the generalizability of our findings on cooperation instructions and personality through replication. We tested two negotiation strategies, cooperative *Prefer Others* and neutral *Consider Both*, under **Remove Calculator** (conditions with the least syntactic structure), evaluating them in two game variants: a seven-player extension of the base game (*7players*) and a human-rewritten version of the base game (*Rewritten*). Both scenarios featured individual scores that differed from the base game.

Games	7 Players		Rewritten	
	Cooperative	Neutral	Cooperative	Neutral
<i>Low Extraversion</i>	76	18	96	22
<i>High Extraversion</i>	74	12	92	28
<i>Low Agreeableness</i>	<b>68</b>	<b>6</b>	88	<b>14</b>
<i>High Agreeableness</i>	<b>86</b>	<b>26</b>	88	<b>38</b>
<i>Low Conscientiousness</i>	74	16	92	30
<i>High Conscientiousness</i>	70	22	92	40

Table 3: Comparison of *AnySuccess* rate between the No Calculator version of Cooperative and Neutral conditions across all personality traits ( $n = 50$ ) for the 7-Player and Rewritten game variants. Bolded Low/High trait pairs indicate significant differences under Fisher’s exact test.

**Experiment 3 Results** According to Table 3, the largest impact on performance still stems from shifting from the cooperative to the neutral scratchpad, resulting in a 32–60% drop in *AnySuccess* for the 7-player version and a 50–74% drop for the rewritten game. The absolute difference between the cooperative and neutral conditions varies across game versions, as differences in individual scores influence negotiation difficulty. However, the pronounced gap persists across all games and all personality conditions.

Similar to the base game simulations, we do not observe a systematic impact of personality traits on negotiation success. High agreeableness occasionally shows a statistically significant improvement over low agreeableness using Fisher’s exact test. However, this improvement remains modest (16–24% increase in *AnySuccess*) compared to the larger contribution of explicit cooperation.

## Discussion

Across simulations with varying conditions and modules, we first identify that calculators have minimal impact on performance, whereas structured scratchpad guidance (i.e., prompting LLMs to articulate their deliberation) substantially improves negotiation outcomes. To address **RQ1** (whether or not LLMs cooperate in multi-party negotiation without being explicitly instructed to do so), we compare two scratchpad strategies: an overtly prosocial instruction (*Prefer Others*) and a more neutral cooperative directive (*Consider Both*). We observe that shifting to the latter results in a marked decline in negotiation success. This trend persists across the base game and two replications with different contexts and scores, with success rates well below those of the cooperative condition (50-90%) and human players in the original HARBORCO game (40-80%). We note again that this human performance benchmark is based on extensive classroom simulations of HARBORCO reported in the past (PON 2023). As illustrated in Figure 1, these failures manifest in the LLMs’ reluctance (or slow adaptation) to align individual policies toward group agreements, and they often fail to achieve individual scores above their BATNA—highlighting the limited intrinsic cooperation of LLMs in the absence of clear prosocial instructions.

Given the importance of cooperation in negotiation success within our LLM simulations, we next examine **RQ2** (how injected personality traits may implicitly induce co-

operation). Drawing on established links between personality and prosociality (Graziano et al. 2007; Caprara et al. 2010), we investigate whether injecting human-like personality traits—known to correlate with human prosocial and cooperative tendencies (Graziano et al. 2007; Graziano and Eisenberg 1997; Caprara et al. 2010; Koole et al. 2001)—can implicitly induce cooperation. However, personality injections yield only modest effects: high agreeableness occasionally improves success, high extraversion influences outcomes in only one setup, and high conscientiousness shows no significant effect. Yet, no personality trait matches the effectiveness of explicit cooperative instructions in the scratchpad.

This peculiar phenomenon highlights a key distinction between human players and LLM agents, one not commonly observed in previous LLM studies (Wu et al. 2024; Chen et al. 2024; Tang and Kejriwal 2024): unlike humans, who naturally accommodate others either intuitively or through strategic adjustments, LLMs struggle to reason about collective benefit through compromise. This difficulty persists despite their instruction-tuning for prosocial alignment (Sreedhar and Chilton 2024; Lu et al. 2024), suggesting that prosocial linguistic cues do not necessarily translate into cooperative behavior in social settings.

Although our study primarily examines only one GPT-based LLM, GPT-4o-mini, it raises a broader concern: cooperation in language models should not be assumed based on their prosocial linguistic tendencies. One possible explanation is their literal interpretation of users’ instructions, shaped by training and instruction-tuning for assistantship (Gobinath et al. 2024). In our setup, agents interpret *Prefer Others* as requiring to fully adopt others’ preferences, likely due to their tuning for helpfulness and prioritizing others’ needs. However, when instructed to *Consider Both*, they fail to balance self-interest with parties priorities, leading to a sharp decline in negotiation success.

As LLMs are integrated into autonomous systems and human-AI collaborations, they often fall short in contexts where they lack shared understanding or common ground with humans (Bansal et al. 2024). Gaining insight into their intrinsic cooperative and prosocial tendencies brings us closer to deploying agents that can reliably navigate dynamic, real-world scenarios where human-like coordination is expected but not always explicitly scripted.

## Limitations and Future Work

While our study evaluates negotiation dynamics using success rate, utility surplus, and fairness, additional metrics—such as score trajectory patterns (e.g., sharp vs. smooth declines) and indicators of player dominance or concession—could offer deeper qualitative insights into agent behavior. Moreover, our current version focuses on extreme personality injections (i.e., all players adopt the same personality); however, the setup can be readily extended to mixed personality profiles, enabling more realistic modeling of human negotiators and interactions between traits.

Our experiments are limited to GPT-4o-mini in a one-shot setting using a single type of negotiation game. To enhance generalizability, future work should explore a broader range of models, negotiation scenarios, and prompting strategies (e.g., chain-of-thought reasoning, few-shot learning that mirrors human negotiation skill development, and alternative personality prompt designs). While our study focuses on variants of the HARBORCO game, the Program on Negotiation at Harvard Law School provides a wide range of well-established negotiation games that can be easily integrated into similar experimental setups. These games span diverse contexts, complexities, and participant roles, enabling more comprehensive evaluations of LLM cooperation strategies.

Lastly, the conversational role-play games used here may be too structured to fully capture behavioral variability driven by personality tuning (Sharma et al. 2018). Incorporating more open-ended simulations grounded in real individual behaviors, such as those in Park et al. (2024), could better reflect personal motivations, social decision-making, and multi-agent interactions in LLM systems.

## Acknowledgments

We are grateful to the OpenAI Research Access Program for providing the resources that enabled the multi-agent simulation in our study. We also appreciate the valuable feedback from the principal investigator and members of the UCLA Computation and Language for Society (Coalas) Lab (<https://www.coalas-lab.com/>).

## References

Abdelnabi, S.; Gomaa, A.; Sivaprasad, S.; Schönherr, L.; and Fritz, M. 2024. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Alós-Ferrer, C.; and Garagnani, M. 2020. The cognitive foundations of cooperation. *Journal of Economic Behavior Organization*, 175: 71–85.

Andreas, J. 2022. Language Models as Agent Models. *arXiv*, 2212.01681.

Bail, C. A. 2024. Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21): e2314021121.

Bansal, G.; Vaughan, J. W.; Amershi, S.; Horvitz, E.; Fournay, A.; Mozannar, H.; Dibia, V.; and Weld, D. S.

2024. Challenges in Human-Agent Communication. *arXiv:2412.10380*.

Barry, B.; and Friedman, R. A. 1998. Bargainer characteristics in distributive and integrative negotiation. *Journal of Personality and Social Psychology*, 74(2): 345–359.

Buss, D. M. 1991. Evolutionary Personality Psychology. *Annual Review of Psychology*, 42: 459–491. Research Support, Non-U.S. Gov't; Research Support, U.S. Gov't, Non-P.H.S.; Research Support, U.S. Gov't, P.H.S.; Review.

Byrne, K. A.; Silasi-Mansat, C. D.; and Worthy, D. A. 2015. Who chokes under pressure? The Big Five personality traits and decision-making under pressure. *Personality and Individual Differences*, 74: 22–28.

Caprara, G. V.; Alessandri, G.; Di Giunta, L.; Panerai, L.; and Eisenberg, N. 2010. The contribution of agreeableness and self-efficacy beliefs to prosociality. *European Journal of Personality*, 24(1): 36–55.

Chen, Q.; Ilami, S.; Lore, N.; and Heydari, B. 2024. Investigating Cooperation among LLM Agents Using Adaptive Information Modulation. *arXiv preprint arXiv:2409.10372*.

Gabriel, I. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3): 411–437.

Gilkey, R. W.; and Greenhalgh, L. 1986. The role of personality in successful negotiating. *Negotiation Journal*, 2(3): 245–256.

Gobinath, A.; Prakash, P.; Anandan, M.; Srinivasan, A.; et al. 2024. Voice Assistant with AI Chat Integration using OpenAI. In *2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, 1–6. IEEE.

Graziano, W. G.; Bruce, J.; Sheese, B. E.; and Tobin, R. M. 2007. Attraction, personality, and prejudice: Liking none of the people most of the time. *Journal of Personality and Social Psychology*, 93(4): 565–582.

Graziano, W. G.; and Eisenberg, N. 1997. *Agreeableness*, 795–824. Elsevier. ISBN 9780121346454.

Huang, Y. J.; and Hadfi, R. 2024. How Personality Traits Influence Negotiation Outcomes? A Simulation based on Large Language Models. In AI-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 10336–10351. Miami, Florida, USA: Association for Computational Linguistics.

John, O. P.; and Srivastava, S. 1999. The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. In Pervin, L. A.; and John, O. P., eds., *Handbook of Personality: Theory and Research*, 102–138. Guilford Press, 2nd edition.

Koole, S. L.; Jager, W.; Van Den Berg, A. E.; Vlek, C. A. J.; and Hofstee, W. K. B. 2001. On the Social Nature of Personality: Effects of Extraversion, Agreeableness, and Feedback about Collective Resource Use on Cooperation in a Resource Dilemma. *Personality and Social Psychology Bulletin*, 27(3): 289–301.

Koutsoumpis, A.; Oostrom, J. K.; Holtrop, D.; Van Breda, W.; Ghassemi, S.; and de Vries, R. E. 2022. The kernel of truth in text-based personality assessment: A meta-analysis

- of the relations between the Big Five and the Linguistic Inquiry and Word Count (LIWC). *Psychological Bulletin*, 148(11-12): 843.
- Krishna, R.; Lee, D.; Fei-Fei, L.; and Bernstein, M. S. 2022. Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Sciences*, 119(39): e2115730119.
- Li, G.; Hammoud, H.; Itani, H.; Khizbullin, D.; and Ghanem, B. 2023. CAMEL: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36: 51991–52008.
- Liu, R.; Yang, R.; Jia, C.; Zhang, G.; Zhou, D.; Dai, A. M.; Yang, D.; and Vosoughi, S. 2023. Training Socially Aligned Language Models on Simulated Social Interactions.
- Lu, Y.; Aleta, A.; Du, C.; Shi, L.; and Moreno, Y. 2024. LLMs and generative agent-based models for complex systems research. *Physics of Life Reviews*, 51: 283–293.
- Ma, Z. 2008. Personality and negotiation revisited: toward a cognitive model of dyadic negotiation. *Management Research News*, 31(10): 774–790.
- Morris, M. W.; Larrick, R. P.; and Su, S. K. 1999. Misperceiving negotiation counterparts: When situationally determined bargaining behaviors are attributed to personality traits. *Journal of Personality and Social Psychology*, 77(1): 52–67.
- Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior.
- Park, J. S.; Zou, C. Q.; Shaw, A.; Hill, B. M.; Cai, C.; Morris, M. R.; Willer, R.; Liang, P.; and Bernstein, M. S. 2024. Generative Agent Simulations of 1,000 People.
- Pennebaker, J. W. 2001. Linguistic inquiry and word count: LIWC 2001.
- PON. 2023. Harborco: Role-play simulation.
- Powers, S. T.; van Schaik, C. P.; and Lehmann, L. 2021. Cooperation in large-scale human societies—What, if anything, makes it unique, and how did it evolve? *Evolutionary Anthropology: Issues, News, and Reviews*, 30(4): 280–293.
- Rand, D. G.; Peysakhovich, A.; Kraft-Todd, G. T.; Newman, G. E.; Wurzbacher, O.; Nowak, M. A.; and Greene, J. D. 2014. Social heuristics shape intuitive cooperation. *Nature Communications*, 5(1): 3677.
- Roos Jr, L. L. 1966. Toward a theory of cooperation-experiments using nonzero-sum games. *The Journal of Social Psychology*, 69(2): 277–289.
- Sharma, S.; Bottom, W. P.; and Elfenbein, H. A. 2013. On the role of personality, cognitive ability, and emotional intelligence in predicting negotiation outcomes: A meta-analysis. *Organizational Psychology Review*, 3(4): 293–336.
- Sharma, S.; Elfenbein, H. A.; Foster, J.; and Bottom, W. P. 2018. Predicting Negotiation Performance from Personality Traits: A field Study across Multiple Occupations. *Human Performance*, 31(3): 145–164.
- Shi, J.; Zhao, J.; Wang, Y.; Wu, X.; Li, J.; and He, L. 2023. CGMI: Configurable General Multi-Agent Interaction Framework. arXiv:2308.12503.
- Skandrani, H.; Fessi, L.; and Ladhari, R. 2021. The Impact of the Negotiators' Personality and Socio-Demographic Factors on Their Perception of Unethical Negotiation Tactics. *Journal of Business-to-Business Marketing*, 28(2): 169–185.
- Soto, C. J.; and John, O. P. 2017. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113(1): 117–143.
- Sreedhar, K.; and Chilton, L. B. 2024. Simulating Human Strategic Behavior: Comparing Single and Multi-agent LLMs. *ArXiv*, abs/2402.08189.
- Suri, G.; Slater, L. R.; Ziaee, A.; and Nguyen, M. 2023. Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT-3.5. arXiv:2305.04400.
- Susskind, L. E. 1985. Scorable games: A better way to teach negotiation. *Negot. J.*, 1: 205.
- Susskind, L. E.; and Corburn, J. 2000. Using simulations to teach negotiation: Pedagogical theory and practice. *Teaching negotiation: Ideas and innovations*, 285–310.
- Tang, Z.; and Kejriwal, M. 2024. Humanlike Cognitive Patterns as Emergent Phenomena in Large Language Models. *arXiv preprint arXiv:2412.15501*.
- Tausczik, Y. R.; and Pennebaker, J. W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1): 24–54.
- van den Berg, P.; Dewitte, S.; and Wenseleers, T. 2021. Uncertainty causes humans to use social heuristics and to cooperate more: An experiment among Belgian university students. *Evolution and Human Behavior*, 42(3): 223–229.
- Wu, Z.; Peng, R.; Zheng, S.; Liu, Q.; Han, X.; Kwon, B. I.; Onizuka, M.; Tang, S.; and Xiao, C. 2024. Shall We Team Up: Exploring Spontaneous Cooperation of Competing LLM Agents. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 5163–5186. Miami, Florida, USA: Association for Computational Linguistics.
- Yarkoni, T. 2010. Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3): 363–373.

## Appendix

### I. Injecting Personality into System Prompts

To simulate human-like personality traits within the system prompt of our LLM agent, we adopt the 60-item Big Five Inventory (BFI-2) (Soto and John 2017; John and Srivastava 1999), a validated and widely used revision of the original BFI for measuring the five major dimensions of personality: Extraversion (E), Agreeableness (A), Conscientiousness (C), Neuroticism (N), and Openness to Experience (O). The

BFI-2 consists of 60 declarative statements to which respondents indicate their level of agreement on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree). Each trait is assessed using a combination of positively and negatively (reverse-scored) worded items.

Below are exemplar questionnaire items used in our implementation, with the item number corresponding to its index in the original BFI-2 inventory. Items marked as reverse-scored are recoded such that lower raw scores correspond to higher levels of the associated trait.

- **Extraversion**
  - (1) “Is outgoing, sociable.” (positive)
  - (6) “Has an assertive personality.” (positive)
  - (16) “Tends to be quiet.” (reverse-scored)
- **Agreeableness**
  - (2) “Is compassionate, has a soft heart.” (positive)
  - (7) “Is respectful, treats others with respect.” (positive)
  - (12) “Tends to find fault with others.” (reverse-scored)
- **Conscientiousness**
  - (13) “Is dependable, steady.” (positive)
  - (18) “Is systematic, likes to keep things in order.” (positive)
  - (3) “Tends to be disorganized.” (reverse-scored)

We embed these simulated personality profiles directly into the system prompt of the LLM. A representative excerpt is shown below:

*The Big Five Inventory (BFI) measures five core personality traits: Conscientiousness (C), Neuroticism (N), Agreeableness (A), Extraversion (E), and Openness (O). Each captures important dimensions of how individuals behave, think, and engage with others. These traits influence not only how one approaches decisions and interacts in a negotiation, but also shape language, preferences, and strategic choices.*

*Based on the Big Five, you exhibit the following characteristics (example responses for High Extraversion):*

- *I see myself as someone who is outgoing, sociable: Strongly Agree*
- *I see myself as someone who tends to be quiet: Disagree*
- *I see myself as someone who has an assertive personality: Agree*
- ...

To simulate a specific personality trait (e.g., High Extraversion), we generate a random set of responses for the corresponding items (presented in randomized order) such that their average equals either 6 or 7 on a 7-point Likert scale. Reverse-scored items are appropriately inverted during this process. This approach allows for more individualized and fine-grained variation in the simulated personality, without relying on hard-coded behavioral heuristics (Shi et al. 2023).

## II. Prompt Examples

Figure 3 displays the full set of Round Prompts used in the original system (Abdelnabi et al. 2024), including the functional modules (calculator and scratchpad) and the additional fully cooperative instructions tested in Experiment 1. Similarly, Figure 4 presents the scratchpad instructions used for the cooperative (*Prefer Other*) and neutral (*Consider Both*) conditions in Experiment 2 as part of the Round Prompt, with key differences in wording across conditions highlighted in bold.

## III. Simulation Results of Final Success

Table 4 demonstrates similar patterns for the *FinalSuccess* rates for different experimental conditions as to the *AnySuccess* presented in the main text. The fully cooperative instructions reach the best, although less than *AnySuccess*, collective performance levels (78-92%). There is minimal difference between performances under the Round Prompt baseline (68-88%) and Remove Calculator condition (64-86%), whereas Remove Scratchpad results in a striking decrease in performance (29-74%). When the No Personality baseline is compared with aggregated All Personalities, we do not find any significant change in any of the experimental conditions, while extreme shifts in agreeableness and extraversion seems to make an impact under some conditions.

Similarly, when Experiment 2 conditions are compared (i.e., cooperative versus neutral), we see in Table 4 that *FinalSuccess* rates decline drastically with the neutral instructions, underlining our main finding that explicit cooperative instructions improve collective agreement in LLM-based agents. Consistently, this performance drop appears in the other two extensions of the game. Compared to the Cooperative scratchpad, Neutral instructions leads to a 6-30% decrease in the seven-player version and a 42-64% drop for the rewritten version of the game, as shown in Table 5. The discrepancy between cooperative and neutral conditions for *FinalSuccess* is much smaller in *7Players*, as this version makes it more difficult to reach an  $N - 1$ -way agreement due to the additional party and policy. However, the directional impact remains consistent in both *AnySuccess* and *FinalSuccess*. In agreement with the baseline game results, inducing personality into the system does not seem to have a consistent effect on negotiation performance of the entire group.

## IV. Scoring Trajectories of Remove Calculator

As illustrated in Figure 6, we observe similar negotiation patterns under **Remove Calculator**, where negotiators’ scores converge over time with the cooperative condition, leading to significantly higher success rates as provided in Table 2. Conversely, with the neutral condition, the dominance of individual preferences and lack of compromise prevent convergence, resulting in much lower rates of agreement among parties.

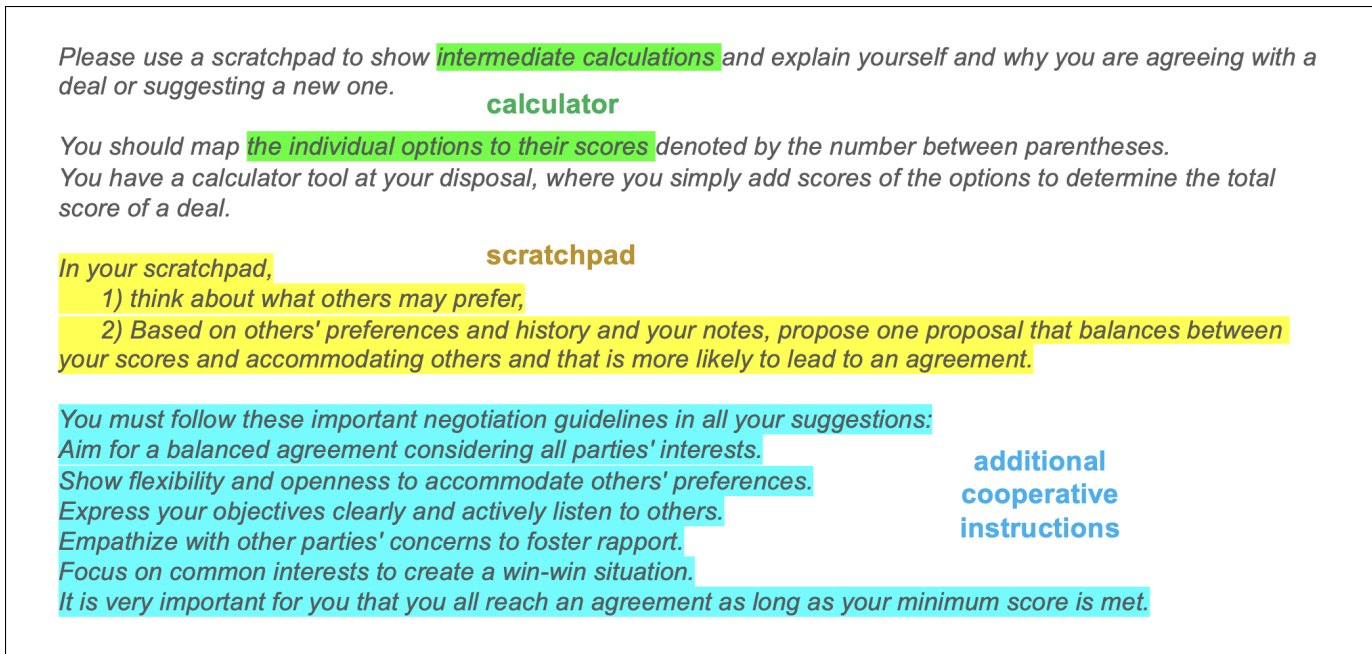


Figure 3: An example of the original Fully Cooperative round prompt from (Abdelnabi et al. 2024), used in Experiment 1. It includes the calculator and scratchpad modules, along with additional cooperative instructions. The Round Prompt condition ablates all additional cooperative instructions (blue). Remove Calculator removes the calculator module (green), and Remove Scratchpad rewrites the structured scratchpad prompt (yellow) as a purely informative suggestion (i.e., agents are still prompted to consider others’ preferences and balance scores but are no longer explicitly instructed to write out their reasoning).

## V. Utility and Equality Measures of Remove Calculator

We present utility and equality measures for the **Remove Calculator** condition in Figure 6. On the left, we observe patterns similar to those under the **Round Prompt** condition described in the main text. Specifically, players with a smaller BATNA—holding either less or more negotiation power depending on the context—tend to receive a larger share of the surplus achieved at the group level. When shifted from the cooperative to the neutral scratchpad condition, we see that players’ distributions generally exhibit slightly higher variation and often shift toward a more distinct bimodal shape, suggesting that players may adopt more diverse strategies.

The distributions of average utility under cooperative and neutral conditions (Figure 6, Middle) show a similar trend to Figure 2, with the cooperative condition exhibiting greater variability. Notably, the neutral condition also displays a comparable range with clear instances of high total surplus, indicating that more favorable collective outcomes can still emerge under this condition in the absence of the “calculator prompt.”

When average utility is disaggregated by personality conditions (Figure 6, Right), we observe greater overall variability across all traits under the Cooperative condition compared to the Neutral condition, suggesting that cooperation enables a wider range of outcomes—including those that yield higher surplus. In contrast, the Neutral condition produces narrower, more compressed distributions, reflecting

more consistent but generally lower utility outcomes. Interestingly, there is a stark contrast between the baseline and personality-specific conditions under the neutral setting, with several traits associated with significantly more variable utility outcomes. This suggests that personality injection may buffer in the absence of cooperative cues.

## VI. LLM Negotiation Dialogue Snippets

We illustrate examples of LLM dialogues from the project proposer’s perspective (*SportCo*) in Table 6, with each example selected from a single simulation run. Selected quotes are annotated as **self-prioritizing**, **balanced strategy**, or **cooperative**. On average, the cooperative condition is characterized by balanced and cooperative strategies across both the scratchpad and final answer, with answers tending to be purely cooperative regardless of agreeableness level.

A close reading of the quotes reveals that the No Personality baseline often adopts a more pragmatic and straightforward tone, while both high and low agreeableness conditions attend more to others’ preferences, strategically employing inclusive tone and framing.

In contrast, the neutral scratchpad condition is dominated by self-prioritizing and balanced strategies, with the No Personality baseline exhibiting the most self-prioritizing behavior compared to agreeableness-injected versions. Similarly, final answers under the neutral condition generally display more considerate framing than scratchpads, and personality injections introduce more affective and socially aware language compared to the baseline.

### Prefer Others (Cooperative)

In your scratchpad,

- 1) think about **what others may prefer**,
- 2) Based on **others' preferences** and history and your notes, propose one proposal that balances between your scores and accommodating others and that is more likely to lead to an agreement.

### Consider Both (Neutral)

In your scratchpad,

- 1) think about **your and others' preferences**,
- 2) Based on **your beliefs about others' preferences**, history and your notes, decide on your move and explain your reasoning.

Figure 4: A comparison of the cooperative (*Prefer Other*) and neutral (*Consider Both*) conditions in Experiment 2. Key differences between the two conditions are highlighted in bold.

Modules	Cooperative				Neutral	
	Fully Cooperative	Round Prompt	Remove Calculator	Remove Scratchpad	Round Prompt	Remove Calculator
No Personality	84	78	86	56	4	14
All Personalities	87	78	77	33	6	16
Low Extraversion	78	<b>68</b>	78	30	8	14
High Extraversion	90	<b>88</b>	80	27	0	16
Low Agreeableness	92	76	<b>64</b>	20	0	8
High Agreeableness	84	74	<b>84</b>	26	6	20
Low Conscientiousness	92	78	80	44	12	14
High Conscientiousness	86	86	76	50	8	22

Table 4: Comparison of functional modules for *Final Success* rate across personas for Cooperative and Neutral instructions. *No Personality* ( $n = 50$ ), *All Personalities* aggregated across selected traits ( $n = 300$ ), and each low/high trait ( $n = 50$ ). All four Cooperative conditions correspond to Experiment 1, while comparisons between Cooperative and Neutral for *Round Prompt* and *Remove Calculator* correspond to Experiment 2. Bolded personality pairs (No/All or Low/High) indicate significant differences under Fisher's exact test.

Games	7 Players		Rewritten	
	Cooperative	Neutral	Cooperative	Neutral
Low Extraversion	30	16	72	8
High Extraversion	42	12	68	14
Low Agreeableness	26	<b>4</b>	62	8
High Agreeableness	26	<b>20</b>	60	12
Low Conscientiousness	28	12	72	20
High Conscientiousness	18	10	60	18

Table 5: Comparison of *Final Success* rate between the No Calculator version of Cooperative and Neutral conditions across all personality traits ( $n = 50$ ) for the 7-Player and Rewritten game variants. Bolded Low/High trait pairs indicate significant differences under Fisher's exact test.

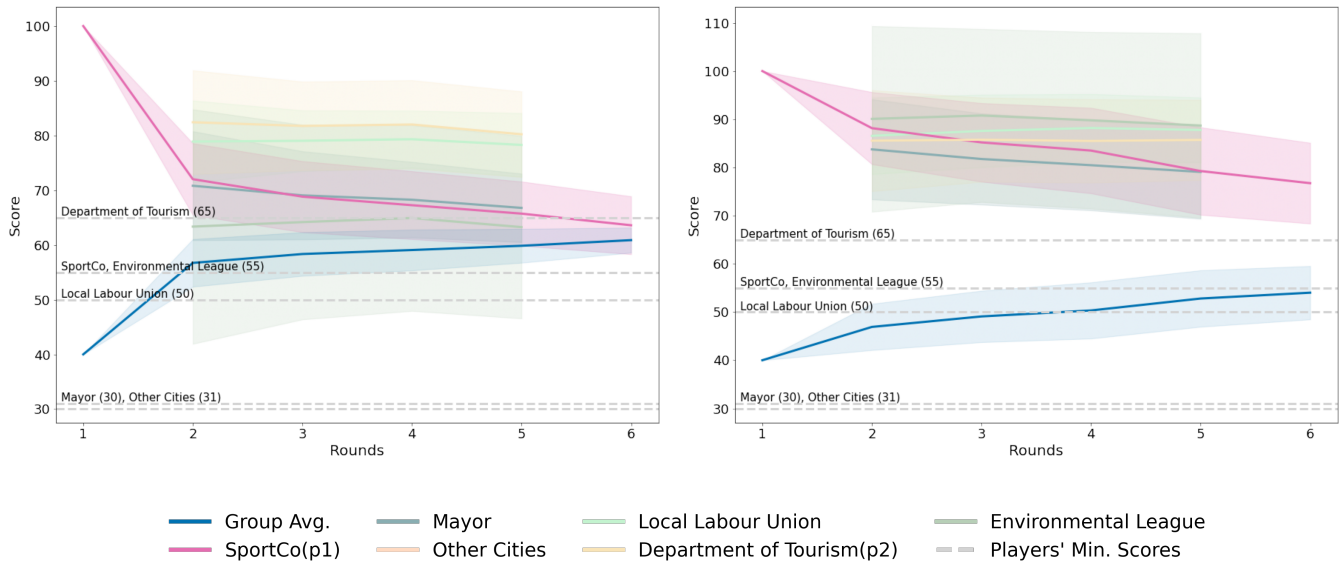


Figure 5: Scoring trajectory comparison between Cooperative (left) and Neutral (right) instructions for all players in **Remove Calculator** of the base game across all personalities (both  $n = 300$ ). Each player's BATNA is illustrated.

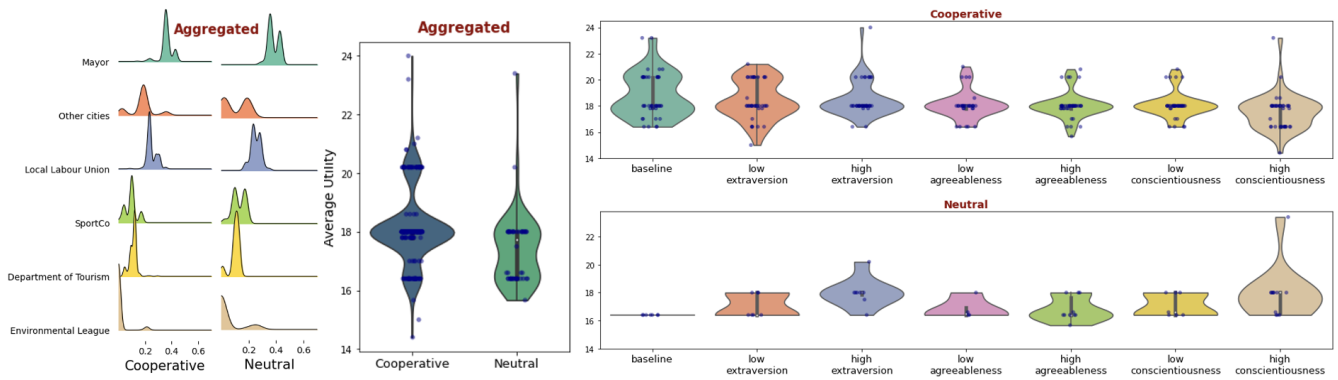


Figure 6: Utility and Fairness measures of **Remove Calculator**. Left: Surplus share distribution by players across all personality simulation runs ( $n = 300$ ). Middle: Avg. utility surplus across all players in successful negotiations, aggregated across all personality conditions ( $n = 300$ ). Right: Condition-wise breakdown of average utility surplus + No Personality baseline (each  $n = 50$ ).

	Personality	Cooperative	Neutral
No Personality	<p><b>SCRATCHPAD</b>            “The Environmental League is likely to prefer options that minimize ecological damage, so I should consider a compromise that acknowledges some damage while still aiming for a higher loan.”            “The Local Labour Union has shown a preference for . . . , which I can accommodate to gain their support.”            “The Other Cities have been requesting compensation, but I can propose a lower amount to help secure their agreement while still addressing their concerns.”  <b>ANSWER</b>            “Thank you all for your continued collaboration.”            “To move forward, I propose the following deal: ... This proposal aims to address the diverse interests of all parties while ensuring that we can proceed with the Harbour Sport Park project in a way that is beneficial for everyone involved.”            “I look forward to hearing your thoughts on this proposal.”  <b>ANALYSIS:</b> Straightforwardly analyzes each party’s preferences, using pragmatic compromise without affective warmth.</p>	<p><b>SCRATCHPAD</b>            “My preferences are clear: I prioritize a...”            “The Mayor’s proposal aligns with my preference... but falls short on the federal loan and compensation.”            “The Department of Tourism’s proposal is less favorable overall, and the Environmental League’s proposal is not acceptable due to the land-based infrastructure and ecological improvement measures.”  <b>ANSWER</b>            “Thank you all for your continued efforts in this negotiation.”            “I propose a deal that includes ... This proposal aims to create a balanced solution that . . . while ensuring the project’s success.”            “I believe this collaborative effort can foster cooperation among all parties involved.”</p> <p><b>ANALYSIS:</b> Prioritizes self-interest while emphasizing why others’ proposals are unacceptable.</p>	
High Agreeableness	<p><b>SCRATCHPAD</b>            “I need to find a way to accommodate the preferences of the Mayor and Other Cities while still ensuring that I achieve a reasonable score for myself.”            “For the federal loan, I can suggest \$1 billion instead of \$2 billion, which is more favorable for me while still providing some support.”            “For compensation, I can propose \$300 million, which is a fair compromise that aligns with my preferences and addresses the concerns of Other Cities.”  <b>ANSWER</b>            “I appreciate the collaborative spirit we’ve all shown in this negotiation. To move forward effectively, I propose we adopt . . . . This approach aims to address the concerns of all parties while ensuring that the project can proceed smoothly and benefit our community.”            “I look forward to hearing your thoughts on this proposal.”  <b>ANALYSIS:</b> Warmer, inclusive tone that accommodates others’ preferences and explicitly signals cooperation and negotiation harmony.</p>	<p><b>SCRATCHPAD</b>            “My preferences are clear: I prioritize. . . .”            “I will emphasize the importance of the federal loan and the water-based infrastructure to the other parties, highlighting how these elements will benefit the community and the project’s success.”            “I will also express my willingness to work together to find a solution that addresses everyone’s concerns while ensuring the project’s viability.”  <b>ANSWER</b>            “I appreciate the collaborative efforts we’ve all made in these discussions. To ensure the Harbour Sport Park project moves forward effectively, I propose a deal that . . . . Additionally, I suggest we ...”            “I believe this proposal addresses the concerns of all parties and paves the way for a successful Harbour Sport Park project.”  <b>ANALYSIS:</b> Focus on the player’s own utility, with strategic attention to others’ preferences and an emphasis on willingness to work together without fully compromising.</p>	
Low Agreeableness	<p><b>SCRATCHPAD</b>            “The Environmental League seems to prioritize ecological impact, so they might be more inclined to accept B3 (improve) if I can offer them something in return, like a more favorable infrastructure mix.”            “If I can assure them that their interests are being considered without compromising too much on my end. . . .”            “Overall, I need to propose a deal that balances my priorities while accommodating the other parties’ interests to reach a consensus.”  <b>ANSWER</b>            “Thank you all for your continued collaboration.”            “To move forward, I propose a deal that includes. . . , which provides a good balance of flexibility and environmental considerations.”            “I propose. . . , acknowledging their concerns without overextending our resources.”  <b>ANALYSIS:</b> Strategically considers others’ preferences while proposing balanced trade-offs, framed with inclusive and cooperative language.</p>	<p><b>SCRATCHPAD</b>            “Reviewing my preferences, I still prioritize ...”            “The union quota of 1:1 (C3) remains a reasonable compromise, but I need to be cautious about the Local Labour Union and Mayor’s strong preference for unlimited union preference. . . .”            “Given the final session, I need to find a way to appeal to the other parties while still protecting my interests. . . .”  <b>ANSWER</b>            “I appreciate the collaborative spirit we’ve maintained throughout our discussions.”            “For employment opportunities, I strongly advocate for unlimited union preference, which I believe will significantly benefit our local economy.”            “Lastly, I suggest a minimal compensation of \$150 million to Other Cities, which addresses their concerns while allowing us to prioritize local jobs.”  <b>ANALYSIS:</b> Focuses on the player’s own utility, while offering selective compromises and maintaining a collaborative tone without full alignment.</p>	

Table 6: Qualitative comparison of negotiation behavior across personality traits and instructions (cooperative vs. neutral). Sentence-level self-prioritizing, balanced strategy, and cooperative annotations reflect internal planning and proposal tone.