Can LLMs Truly Embody Human Personality? Analyzing AI and Human Behavior Alignment in Dispute Resolution

Deuksin Kwon[1,2*], **Kaleen Shrestha**[1*], **Bin Han**[1,2], **Spencer Lin**[1], **James Hale**[1,2], **Jonathan Gratch**[1,2], **Maja Matarić**[1], **Gale M. Lucas**[1,2]

[1]Department of Computer Science, University of Southern California
[2]USC Institute for Creative Technologies
{deuksink, kshresth, binhan, linspenc, jahale, mataric}@usc.edu
{gratch, lucas}@ict.usc.edu

Abstract

Large language models (LLMs) are increasingly used to simulate human behavior in social settings such as legal mediation, negotiation, and dispute resolution. However, it remains unclear whether these simulations reproduce the personality–behavior patterns observed in humans. Human personality, for instance, shapes how individuals navigate social interactions, including strategic choices and behaviors in emotionally charged interactions. This raises the question: *Can LLMs, when prompted with personality traits, reproduce personality-driven differences in human conflict behavior?* To explore this, we introduce an evaluation framework that enables direct comparison of human-human and LLM-LLM behaviors in dispute resolution dialogues with respect to Big Five Inventory (BFI) personality traits. This framework provides a set of interpretable metrics related to strategic behavior and conflict outcomes. We additionally contribute a novel dataset creation methodology for LLM dispute resolution dialogues with matched scenarios and personality traits with respect to human conversations. Finally, we demonstrate the use of our evaluation framework with three contemporary closed-source LLMs and show significant divergences in how personality manifests in conflict across different LLMs compared to human data, challenging the assumption that personality-prompted agents can serve as reliable behavioral proxies in socially impactful applications. Our work highlights the need for psychological grounding and validation in AI simulations before real-world use.

Introduction

The ability to simulate human behavior in high-stakes interpersonal contexts—such as negotiation and dispute resolution—has growing societal relevance. As large language models (LLMs) are increasingly deployed in socially impactful settings, from conflict resolution coaching to AI-assisted decision-making (Lin et al. 2024; Ashtiani and Raahemi 2023; Li et al. 2025; Kwon et al. 2025), it is essential to assess whether they reflect core psychological factors that guide human behavior. One such factor is personality, which plays a central role in shaping how people navigate interpersonal conflict (Antonioni 1998; Bell and Blakeney 1977; Wood and Bell 2008).

Conflict resolution often involves dynamic and strategic decisions—such as whether to assert, accommodate, or withdraw—made under conditions of uncertainty and evolving interpersonal dynamics (Ross and Stillinger 1991). Among the various factors influencing these behaviors, personality traits such as agreeableness, neuroticism, and openness to experience systematically predict individual differences in conflict-related behavior (Antonioni 1998; Tehrani and Yamini 2020). Accounting for personality provides important insights into the social-cognitive mechanisms that drive interpersonal dynamics.

Accordingly, a multitude of studies have examined personality's role in conflict resolution; however, most have relied on static questionnaires or simplified decision tasks (Sternberg and Soriano 1984; Wood and Bell 2008), limiting insight into how personality is behaviorally expressed in complex, unfolding, goal-directed interactions (Baumeister, Vohs, and Funder 2007). Furthermore, research has rarely moved beyond human-only paradigms or LLM-only simulations to examine whether personality-linked behaviors remain consistent across systems. Given the growing use of LLMs to simulate social behavior, this raises a critical empirical question: *Can LLMs reproduce personality-driven differences in human conflict behavior when guided by personality prompts?*

While recent studies have begun exploring personality-prompted LLMs for generating individualized dialogue (Serapio-García et al. 2023; Jiang et al. 2023), the assumption that these models serve as psychologically accurate proxies remains largely untested. The behavioral fidelity of LLMs, particularly in emotionally charged settings, has not been rigorously validated. This gap limits our ability to trust LLMs in applications where misalignment with human behavior could have social/ethical consequences.

To address this gap, we introduce an evaluation framework for systematically assessing how personality manifests in human-human and LLM-LLM conflict resolution dialogues. Leveraging the KObe DISpute corpus (KODIS) dataset (Hale et al. 2025), which features multi-issue, turn-based negotiations embedded in disputes, we construct a

*These authors contributed equally.

parallel set of LLM dialogues by prompting LLMs with matched scenarios and Big Five Inventory (BFI) personality profiles. This framework facilitates fine-grained comparisons of strategic behaviors and conflict outcomes between human and LLM interactions.

Methodologically, we introduce an experimental paradigm that mirrors a human–human dialogue corpus by matching agents on personality and fixing scenarios, enabling one-to-one human–LLM comparisons, and we propose a novel, generalizable framework to evaluate LLMs' behavioral alignment with human psychological constructs. Empirically, applying this framework to recent LLMs reveals consistent mismatches: in humans, neuroticism is the strongest predictor of strategic outcomes, whereas models show stronger effects for extraversion and agreeableness and exhibit strategic behaviors that load on a broader set of personality factors. Among LLMs, Claude and Gemini align more closely with human strategic metrics than GPT-4o mini, indicating partial convergence with human patterns. Our framework, simulation code, and supplementary materials are publicly available [1].

These findings challenge the growing assumption that personality-prompted LLMs can reliably serve as human proxies in social applications. As LLMs become more integrated in socially consequential domains, our work underscores the need for psychological grounding, interpretability, and behavioral validation. Ultimately, our framework provides a path toward more human-aligned, socially responsible AI systems.

Background and Related Work

To inform our research, we investigate prior work on personality and conflict, as well as personality prompting of LLMs.

Personality and Conflict Resolution Strategy

Personality traits are influential factors in how individuals perceive and respond to interpersonal conflict (Wood and Bell 2008; Tehrani and Yamini 2020; Antonioni 1998). Early studies identified links between the BFI personality traits and conflict resolution styles. For example, Jones and Melcher (1982) and Bell and Blakeney (1977) found that preferences for competing, accommodating, and avoiding conflict resolution styles vary based on personality. Wood and Bell (2008) further showed that extraversion and agreeableness significantly predict conflict style based on the Thomas–Kilmann model. Participants with higher extraversion or agreeableness are more likely to adopt collaborative or accommodating approaches. Kaushal and Kwantes (2006) examined how personality-related traits, specifically emotional intelligence, self-monitoring, and cultural values, influence conflict resolution strategy. The study found that collectivism and interpersonal harmony were associated with preferences for harmony-preserving strategies. Park and Antonioni (2007) used questionnaires and found that extraversion and agreeableness were linked to cooperative conflict strategies, especially when the other party was cooperative. Neurotic individuals, in contrast, favored avoidant or dominating approaches regardless of the partner's behavior. Graziano, Jensen-Campbell, and Hair (1996) examined how agreeableness shapes conflict resolution preferences and behavior. Agreeable individuals consistently favored negotiation and disengagement over power assertion.

LLMs with Personality

Recent advances in LLMs have enabled the simulation of complex human social behaviors (Zhou et al. 2023; Park et al. 2022). These advances have enabled research into simulations between multiple LLMs, in domains such as negotiation and other decision-making contexts (Kwon et al. 2025; Abdelnabi et al. 2024; Xie et al. 2024).

To enrich social behavior, recent studies have explored prompting LLMs to emulate personality profiles, particularly those based on BFI (Serapio-García et al. 2023; Jiang et al. 2023; Sorokovikova et al. 2024; Gui and Toubia 2023; Han et al. 2025). Incorporating such traits allows LLMs to better reflect personality differences, enriching the psychological realism and diversity of multi-LLM interactions.

Serapio-García et al. (2023) and Jiang et al. (2023) proposed prompting frameworks for eliciting and modeling personality traits in LLMs. Other works show LLMs can exhibit or be edited toward trait-consistent responses (Sorokovikova et al. 2024; Mao et al. 2024). Building on these capabilities, recent work has employed personality-conditioned LLM simulations. Notably, Huang and Hadfi (2024) used BFI-prompted LLMs in negotiation settings and observed trait-driven differences in strategy and preferences. This line of research highlights LLMs' potential for simulating psychologically grounded behavior at scale.

However, much of the past work assumes that LLMs prompted with specific traits behave analogously to humans with similar dispositions. This assumption, however, remains largely untested, especially in complex social contexts such as conflict and negotiation, where behavior reflects internal traits and interpersonal dynamics. Prior studies have focused on trait expression, not on whether LLM behaviors align with human behavior. Our study systematically compares personality–behavior relationships in LLMs and humans under matched negotiation scenarios, highlighting key differences in behavior.

Methodology

To explore personality in conflict resolution dialogue and compare human and LLM behavior, we introduce the datasets and conflict resolution behavioral measures to investigate our proposed research questions.

KODIS: Human vs. Human Dataset

The *KObe DISpute corpus* (KODIS) is a role-play dispute resolution dataset collected from crowd-sourced participants on Prolific by Hale et al. (2025). It consists of extended dialogues between two participants (out of 4,061) with Buyer or Seller roles. As shown in Figure 1, the scenario is an emotionally charged dispute over a jersey purchased online for

[1] https://github.com/DSincerity/Personality-LLM-BehavAlign-Dispute

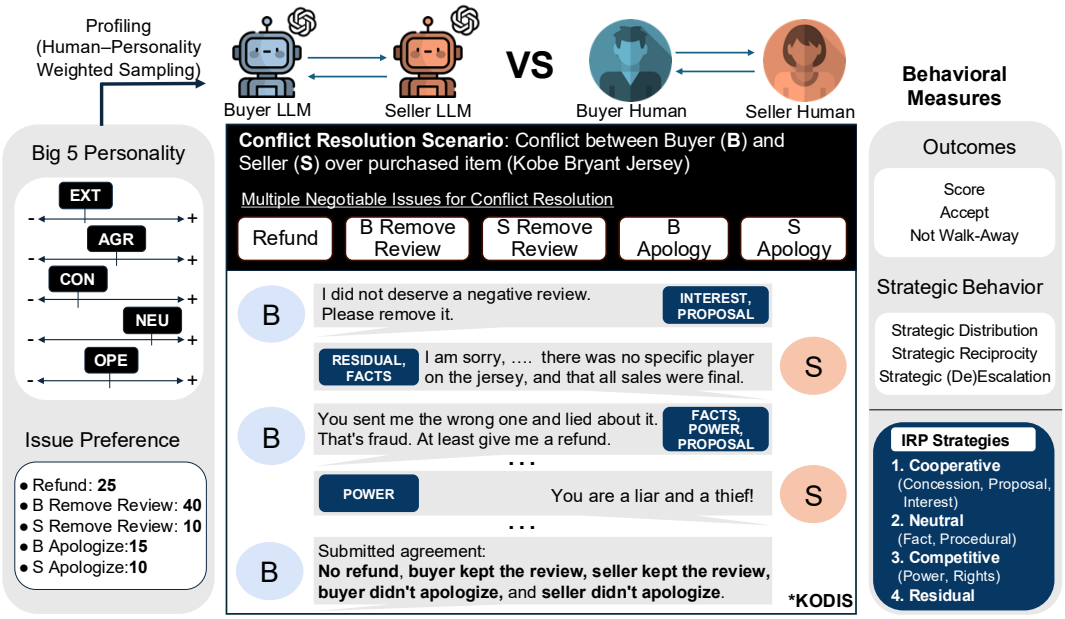

Figure 1: Overview of the conflict resolution scenario, an example dialogue from the KODIS dataset, the profiling setup for LLM simulation, and the behavioral evaluation measures.

a sick nephew. The two parties have different perspectives and strong negative emotions towards each other. They then negotiate a resolution based on their preference over a set number of issues: (1) a full refund, (2) removal of negative reviews, or (3) a formal apology. We excluded human–AI conversations from the original dataset and used a subset of 248 human–human dialogues with complete personality information from both participants. Each analysis used a filtered subset with missing values removed as appropriate.

Behavioral Measures

We define two types of behavioral measures: (1) final outcomes, and (2) strategic behaviors.

Final Outcomes We focus on three final outcome variables of dispute conversation, corresponding to well-established negotiation measures (Kelley 1996).

(1) *Score*: The participant's final payoff, computed as the inner product of the agreed allocation and the participant's issue preferences.

(2) *Accept*: Whether the participant accepted an offer.

(3) *Not Walk-Away*: Whether the participant chose to stay in the negotiation rather than walking away. Staying reflects a basic strategic commitment to reach an agreement.

Strategic Behaviors To investigate strategic behaviors, we present four metrics based on the Interests-Rights-Power (IRP) framework (Ury, Brett, and Goldberg 1988). This framework categorizes utterances according to whether they appeal to interest, assert rights, or leverage power, offering a useful lens for understanding how individuals navigate conflict. Figure 1 illustrates IRP annotations within a conflict dialogue, showing strategy shifts. The IRP framework defines eight utterance types, as shown in Table 1.

In prior work, BFI personality traits have been linked to conflict styles (Tehrani and Yamini 2020). We examine whether participants' use of IRP strategies—particularly the most relevant in conflict, *Cooperative* and *Competitive* strategies (Brett, Shapiro, and Lytle 1998)—is associated with personality. Cooperative mirroring aids resolution (Kelley 1996), while competitive mirroring can push conflict into destructive spirals (Brett, Shapiro, and Lytle 1998). Thus, we analyze whether participants reciprocate their partner's *Cooperative* or *Competitive* IRP strategies in the immediate turn, and how these patterns relate to personality. Additionally, we examine escalation and de-escalation, defined as competitive responses to non-competitive moves (and vice versa). Reciprocity and escalation are well-studied concepts in negotiation and conflict research (Putnam and Jones 1982; Zartman and Faure 2005). Speakers may use these strategically—to assert dominance or to reduce tension and refocus the negotiation (Brett, Shapiro, and Lytle 1998). Accordingly, we define the following four metrics:

(1) **IRP Ratio**: The relative frequency of competitive (C) or cooperative (Co) strategies used by the speaker (S):

$$\text{IRP}^X_{\text{ratio}} = \frac{N_S^X}{N_S^{\text{all strategies}}}, \quad X \in \{C, Co\}$$

(2) **IRP Reciprocity**: The proportion of speaker turns that match the partner's preceding strategy (X_P):

$$\text{IRP}^X_{\text{recip}} = \frac{N_S^{X=X_P}}{N_P^X}, \quad X \in \{C, Co\}$$

(3) **Escalation Ratio**: The rate at which the speaker responds competitively to non-competitive partner turns (NC_P):

$$\text{Escalation} = \frac{N_S^{NC_P \to C}}{N_P^{NC}}$$

Strategy	Examples and Definitions
Cooperative	**Proposal**: Concrete ideas for resolving the conflict; **Concession**: Willingness to change an initial position; **Interests**: Statements of needs or concerns; **Positive Expectations**: Optimistic outlook or recognition of common goals.
Neutral	**Facts**: Information sharing or clarification; **Procedural**: Statements about conversation structure or norms.
Competitive	**Power**: Threats or coercive moves; **Rights**: Appeals to rules, norms, or fairness.
Residual	**Residual**: If an utterance does not fit into any of the above categories.

Table 1: IRP strategy types and example speech acts.

(4) **De-escalation Ratio**: The rate at which the speaker responds non-competitively to competitive partner turns:

$$\text{De-escalation} = \frac{N_S^{C_P \to NC}}{N_P^C}$$

Here, N_S^X denotes the number of speaker turns using strategy X, and $N_S^{Y_P \to X}$ refers to speaker turns using X in response to the partner's preceding strategy Y. We define *C* and *Co* as competitive and cooperative strategies, respectively, and *NC* as non-competitive strategies (i.e., *Cooperative*, *Neutral*, and *Residual*).

L2L: LLM-to-LLM Simulation for Parallel Dataset Construction

We construct diverse personality profiles and generate corresponding prompts to create LLMs with distinct personality traits. We build the LLM-to-LLM (L2L) dataset, consisting of LLM-based simulations, and analyze their behaviors in conflict resolution. We use OpenAI GPT-4o mini (Hurst et al. 2024), Anthropic Claude Sonnet 3.7 (Anthropic 2025), and Gemini 2.0 Flash (Google 2025), hereafter GPT-4, Claude, and Gemini. We obtain IRP annotations for the L2L dataset using the same procedure as KODIS. We use the default hyperparameters for all LLMs (temperature of 1) to test the zero-shot learning capabilities of these models.

LLM Personality Profile Each LLM is assigned a BFI personality profile ($\{P_{AGR}, P_{EXT}, P_{CON}, P_{NEU}, P_{OPE}\}$), following the validated personality-prompt design of Huang and Hadfi (2024). This approach produces results highly consistent with IPIP tests, ensuring reliable personality manipulations. Prompt variation is minimized to assess if LLMs can inherently represent human personality traits.

To enable fair comparisons with human data, personality profiles are sampled from the empirical distribution of human BFI traits. Each trait uses a six-point polarity–degree scale (e.g., $\{P_{\text{AGR}}^{++}, P_{\text{EXT}}^{+++}, \dots\}$).

Building on this, each LLM is assigned a personalized issue importance profile and negotiates more assertively on issues it deems more important. For the *Apology*, importance is weighted by LLM agreeableness based on regression results from human data (B=2.13, p=.02). Other issues are assigned importance values at random. This personality-informed prioritization introduces psychologically meaningful variation for fair comparison with human behavior.

To generate personality prompts, we use 70 pairs of bipolar adjectives empirically associated with the BFI (Goldberg 1992; Serapio-García et al. 2023). Each LLM receives 15 adjectives (three per trait), modified to reflect trait intensity: "very" for high, "a bit" for low, and no modifier for medium. We prompt LLMs with their assigned profile prompt throughout the simulation.

LLM Simulation Using the approach described above, we configure two LLMs—one acting as the Buyer and the other as the Seller—and simulate the conflict scenario from KODIS. As shown in Table 1, they negotiate over three core concerns (refund, negative reviews, apology) and must ultimately reach agreement across five issues.

At each turn, LLMs respond to the opponent's previous utterance by making an offer (`SUBMIT`), responding (`ACCEPT`/`REJECT`), or continuing the dialogue. When negotiation reaches deadlock, LLMs can choose to `WALK-AWAY`.

The dialogue ends when one of the LLMs chooses to `ACCEPT` an offer or walk away. The negotiation is considered unsuccessful if neither LLM reaches an agreement within a predefined length limit (`NO AGREEMENT`).

We first run 500 simulations with GPT-4 and then include additional models (250 simulations for each) to demonstrate the framework's flexibility. While simulation counts varied for practical reasons, each provided statistically valid estimates. The L2L Prompt and example can be found in the supplementary material.

Comparison with Regression Analysis

To examine how personality traits influence negotiation behavior, we treat the above proposed metrics as dependent variables (DVs). Based on each DV's data type—continuous or binary (0 or 1)—we applied linear regression or logistic regression, respectively. As independent variables, we included the LLM's own BFI personality traits and those of the partner. Position (e.g., Buyer or Seller) was included as a control variable, using effect coding (Buyer=–1 and Seller=1) (See the supplementary material for model details). These metrics support systematic comparison between humans and LLMs.

Results and Discussion

This section presents results addressing how personality traits shape outcomes and strategic behavior in human conflict dialogues, and whether personality-prompted LLMs align with human behavior. We analyze findings from the KODIS and L2L datasets. Overall, our results reveal divergences between human and LLM behaviors, challenging as-

Dependent Variables (DVs)	Significant Independent Variables (IVs) - Beta Coefficient			
	GPT-4	Claude	Gemini	KODIS
Score	S-EXT (B=1.67**) S-AGR (B=-4.38***) POS (B=16.85***)	S-AGR (B=-2.50***) P-EXT (B=-1.42*) P-AGR (B=3.05***) POS (B=-12.05***)	S-AGR (B=-4.48***) S-CON (B=1.72*) S-OPE (B=1.94*) POS (B=-5.31***)	POS (B=-3.21*)
Accept	POS (B=-0.22*)	S-EXT (B=-0.17**) P-EXT (B=0.17**) POS (B=-0.47***)		S-NEU (B=-0.26*) P-NEU (B=0.27*) POS (B=0.49***)
Not Walk-Away	S-OPE (B=-0.18*) P-OPE (B=0.18*) POS (B=0.94***)		S-NEU (B=-0.18*) P-NEU (B=0.18*)	
Coefficients (B) are reported. *, **, *** indicate p $<$ .05, .01, and .001, respectively.				

Table 2: Summary of significant regression results for personality predictors across LLM (L2L) and human (KODIS) datasets, for conflict resolution action and outcomes. S and P refer to Self and Partner, respectively.

sumptions of LLM alignment. For all regression findings, we report significant results with $p < .05$.

Effects of Personality Traits on Actions and Outcomes

We examined how personality traits influence three key negotiation outcomes—*score, offer acceptance*, and *not walk-away* behavior—comparing humans (KODIS) with LLMs to evaluate the degree of behavioral alignment (See Table 2). Full results can be found in the supplementary material.

Score Human scores showed no significant associations with any personality trait, suggesting that outcomes were shaped more by interactional dynamics than by stable dispositions. Conversely, all LLMs exhibited clear trait-based effects. Higher agreeableness predicted lower scores. For GPT-4, agreeableness negatively influenced outcomes across both roles (Buyer: B=–2.62, p=.001; Seller: B=–3.32, p=.000). Model-specific patterns also emerged: GPT-4 performed better when more extraverted (B=1.67, p=.005); Gemini benefited from higher conscientiousness and openness; and Claude achieved higher scores when paired with introverted but agreeable partners, reflecting responsiveness to partner traits.

Accept Humans showed role-contingent effects of neuroticism. Individuals high in neuroticism were significantly less likely to accept offers (B=–0.26, p=.026), while negotiating with neurotic partners increased acceptance likelihood (B=0.27, p=.025). This asymmetry suggests reluctance to commit under emotional sensitivity and possible appeasement of emotionally volatile counterparts. Among LLMs, GPT-4 and Gemini showed no significant trait effects, but Claude was more likely to accept when introverted and paired with an extroverted partner—a pattern not mirrored in humans.

Not Walk-Away For humans, no personality traits significantly predicted the likelihood of staying in the negotiation, suggesting that such decisions were driven more by contextual factors. In contrast, GPT-4 and Gemini were significantly less likely to remain engaged when personality mismatches with their partners were high, indicating a unique LLM sensitivity to interpersonal dissimilarity that may lead to premature disengagement under trait misalignment.

Role-Dependent Personality Effects and Human–LLM Alignment We also examined whether personality effects vary by negotiation role (Buyer vs. Seller) and found partial alignment between LLMs and humans in role-contingent behavior. While both GPT-4 and humans showed role-dependent effects of agreeableness in offer acceptance, only LLMs showed personality effects on final scores. Full results can be found in the supplementary material.

Effects of Personality Traits on Strategic Behavior

We examined how personality relates to the frequency and reciprocity of *Competitive* and *Cooperative* IRP strategies, as well as the frequency of escalation and de-escalation responses in the KODIS and L2L datasets (Table 3). Full results can be found in the supplementary material.

Personality Effects on IRP Ratio As presented in Table 3 for L2L, several significant personality effects emerged for the frequency of *Cooperative* and *Competitive* strategies.

For *Cooperative* strategies, GPT-4 showed a significant effect of partner extraversion (B=0.49, p=.031), which did not appear in KODIS. Gemini and Claude showed no significant predictors, similar to KODIS. For *Competitive* strategies, GPT-4 aligned more closely with KODIS, showing significant effects for self- and partner-extraversion (B=0.95, p=.025; B=1.01, p=.017), as well as neuroticism (B=0.88, p=.030; B=0.89, p=.028). Gemini also overlapped with KODIS, with significant effects for both self- and partner-extraversion (B=1.07, p=.043; B=1.09, p=.041). Claude showed no overlap. These results suggest that Gemini better captures personality effects for *Cooperative* and *Competitive* strategy frequencies.

Overall, personality effects on IRP strategy for GPT-4 and Claude show low alignment with KODIS, aside from a partial match observed for Cooperative strategies, whereas Gemini showed a higher overlap for both strategy types.

Figure 2 shows IRP strategy distributions across five personality traits. Trait-driven variation across traits is minimal for both humans and LLMs, suggesting limited sensitivity to personality differences.

Dataset	Significant Independent Variables (IVs) - Beta Coefficient					
	Cooperative Ratio	Competitive Ratio	Cooperative Reciprocity	Competitive Reciprocity	Escalation Ratio	De-escalation Ratio
GPT-4	P-EXT (B=0.49*) POS (B=1.56***)	S-EXT (B=0.95*) S-NEU (B=0.88*) P-EXT (B=1.01*) P-NEU (B=0.89*)	S-AGR (B=1.10***) P-EXT (B=-0.89*) POS (B=-6.58***)	S-EXT (B=1.40*) S-AGR (B=-1.46*) S-NEU (B=1.41*) P-CON (B=-1.41*) POS (B=13.27***)	S-AGR (B=-1.37**) P-EXT (B=1.56**) POS (B=11.21***)	S-AGR (B=1.83***) POS (B=-7.57***)
Gemini	POS (B=1.88***)	S-EXT (B=1.07*) P-EXT (B=1.09*)	S-EXT (B=3.08***) S-AGR (B=1.51**) POS (B=-17.18***)	POS (B=7.75***)	POS (B=3.30**)	S-EXT (B=4.13**) P-EXT (B=2.85*) POS (B=-19.33***)
Claude	POS (B=3.24***)		POS (B=-7.02***)	POS (B=13.20***)	S-AGR (B=-2.45***) S-NEU (B=1.02*) POS (B=6.85**)	
KODIS		S-EXT (B=1.74*) P-EXT (B=1.80*) P-CON (B=-1.77*)	S-OPE (B=4.02*) P-CON (B=-4.98**) POS (B=-3.79*)	POS (B=8.36**)	POS (B=8.25***)	

Coefficients (B) are reported. *, **, *** indicate p $<$ .05, .01, and .001, respectively.

Table 3: Regression results for the IRP-related dependent variables: IRP Ratio, IRP Reciprocity Ratio, and Escalation/De-escalation Ratio. S, P, and POS refer to Self, Partner, and Position, respectively. Full results for individual IRP strategies for ratio and reciprocity metrics can be found in the supplementary material.

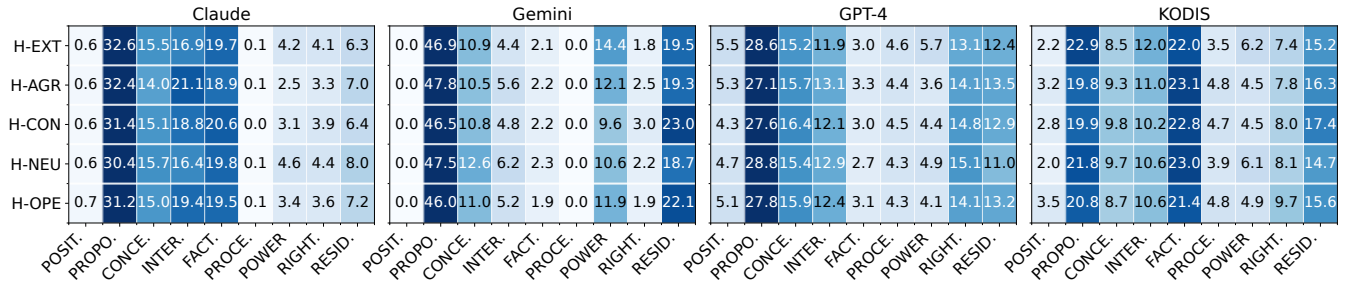

Figure 2: IRP strategy heatmap by personality traits across LLMs and human dialogues; rows sum to 100% (IRP strategy distribution per trait). H denotes "high," and the x-axis labels represent the first five letters of each strategy listed in Table 1.

Humans rely most on *Facts* and show the most balanced distribution, reflecting flexible, context-sensitive behavior. LLMs, in contrast, favor *Proposal* and use *Concession* more, indicating a more transactional style.

Model-specific patterns also vary. Claude most closely resembles humans, with higher use of *Facts*. Gemini shows the most skewed distribution, with elevated *Residual* and *Power*, and no use of *Positive Expectations* or *Procedural* moves. GPT-4 is more balanced but consistently uses more *Power*.

These findings suggest that while LLMs tend to over-rely on transactional strategies, they differ in how rigidly or assertively those strategies are applied, highlighting the need for greater adaptability and nuance.

Personality Effects on IRP Reciprocity For reciprocity of *Cooperative* strategies in the LLMs (Table 3), we find no overlap with the KODIS dataset for GPT-4, with significant effects of partner-extraversion (B=$-0.89, p$=.013) and self-agreeableness (B=$1.10, p$=.000). The same goes for Gemini, where there is also a significant effect of self-agreeableness (B=$1.51, p$=.002), and additionally for self-extraversion (B=$3.08, p$=.000). Claude shows no significant personality effects. For *Competitive* strategies, no overlap with KODIS is observed for GPT-4; in fact, multiple personality traits—self-extraversion, self-agreeableness, self-neuroticism, and partner-conscientiousness—significantly influence reciprocity (B=$1.40, p$=.049; B=$-1.46, p$=.034; B=$1.41, p$=.042; B=$-1.41, p$=.038), whereas there were no significant personality effects in KODIS. Gemini and Claude both align with KODIS, indicating no significant personality effects. These findings demonstrate that Gemini and Claude align with human results for competitive reciprocity, while all three LLMs diverge for cooperative reciprocity, suggesting that competitive reciprocity patterns are more strongly associated with personality.

Personality Effects on (De)Escalation For escalation response frequency, only Gemini aligns with KODIS, with no significant personality effects. GPT-4 and Claude both have significant personality effects, with self-agreeableness in common (B=$-1.37, p$=.004 and B=$-2.45, p$=.000). GPT-4 additionally has significant effects for partner-extraversion (B=$1.56, p$=.002) and Claude additionally has significant effects for self-neuroticism (B=$1.02, p$=.024).

For de-escalation, Claude aligns with KODIS with no significant personality effects. Gemini has significant personality effects for self- and partner-extraversion (B=$4.13, p$=.002) and (B=$2.85, p$=.023). GPT-4 has significant effects for self-agreeableness (B=$-1.83, p$=.000). These results show that Claude was aligned with KODIS on de-escalation, and Gemini was aligned with KODIS on escalation response frequencies, with GPT-4 being least aligned for both measures.

Comparison of IRP Reciprocity and (De)Escalation Frequencies in L2L vs. KODIS Figure 3 compares the frequency of IRP Reciprocity (Cooperative and Competitive)

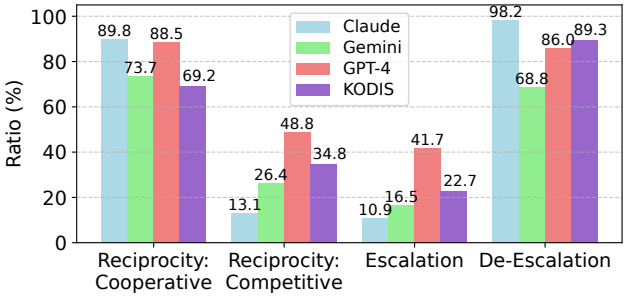

Figure 3: Comparison of frequencies of Cooperative/Competitive reciprocity and (De) Escalation for KODIS and L2L.

and (De) Escalation responses, for humans (KODIS) and LLMs, aggregated across traits due to minimal variation.

LLMs more consistently reciprocate *Cooperative* strategies, whereas humans show more flexible reciprocity, including greater *Competitive* reciprocity. Among LLMs, GPT-4 displays particularly strong responses to competitive moves, suggesting heightened sensitivity to adversarial cues.

Escalation patterns further distinguish models from humans. GPT-4 tends to escalate more readily, while Claude strongly favors de-escalation, with minimal escalation throughout. Humans demonstrate a more balanced use of both, adapting their responses more contextually.

Overall, LLMs reveal more polarized and rigid strategic tendencies, while humans adjust their behaviors more adaptively, highlighting ongoing challenges in behavioral alignment. Personality-specific patterns were similar across models (see supplementary material).

Temporal Dynamics of Strategic Behavior Across Human and LLM Dialogues Figure 4 shows IRP strategy distributions over dialogue stages in high-extraversion cases for humans and LLMs. Full results for other traits can be found in supplementary material. Humans exhibit dynamic progression: *Facts* dominate early stages then decline as *Interest*, *Proposal*, and *Concession* increase, with greater *Residual* use at the end. This reflects structured shifts from factual grounding to relational closure, while *Power* and *Rights* remain minimal. LLMs show flatter trajectories with limited adaptation. Most begin with dominant *Proposal* use, while *Concession* appears earlier and more persistently than in humans, suggesting premature accommodating behavior. Except for Claude, LLMs rarely use *Facts* early on, bypassing the grounding phase. Claude partially mirrors human dynamics with declining *Facts* and increasing *Interest* and *Concession*. Gemini remains *Proposal*-dominated with high *Power* and *Residual* but little temporal variation. GPT-4 shows balanced use but maintains consistently high *Rights*. These patterns reveal that while humans adapt strategies temporally, LLMs follow fixed, model-specific paths, highlighting the need for improved temporal flexibility.

Conclusion

This study provides the first behavior-rich comparison of personality-driven conflict behaviors between humans and LLMs under matched conflict resolution scenarios.

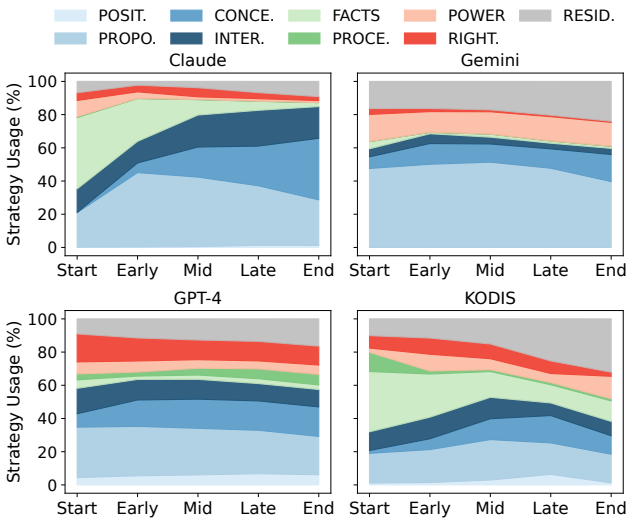

Figure 4: Temporal distribution of the frequency of IRP strategies across dialogue stages in high extraversion cases for human (KODIS) and LLM (L2L) dialog.

Our analyses directly address whether personality-prompted LLMs can replicate human-like behavior in high-stakes interpersonal contexts. While humans flexibly adapt based on both self and partner traits, LLMs show distinct personality-linked patterns that often diverge in how they participate in and resolve conflict, especially in dynamic aspects of resolution rather than static strategy use. These findings highlight key limitations of current prompting approaches and caution against assuming that trait-driven LLMs can reliably proxy human behavior. As LLMs enter socially consequential domains, our results underscore the need for more psychologically grounded and context-sensitive models to ensure safe and responsible deployment.

Limitations and Future Work

This study offers insights into how personality shapes strategic behavior and outcomes, but limitations remain. First, our LLM-based IRP annotations achieved strong F1 scores and partial human validation, but we did not conduct a full-scale review. Second, we minimized prompt variation to test whether LLMs can inherently represent human personality traits, but even small changes in phrasing, instructions, or ordering may shift model behavior; we did not assess robustness to such variations, limiting the generalizability of our findings. Finally, while BFI traits were informative, other personality constructs (e.g., emotional intelligence, Machiavellianism) may offer additional explanatory power.

Future work should incorporate more naturalistic, multimodal conflict data, strengthen annotation validation, and test generalization across LLMs. Our framework could also be extended to analyze linguistic and emotional patterns and to support LLM alignment with personality–behavior relationships observed in human data.

Acknowledgements

Research was sponsored by the Army Research Office under Cooperative Agreement Number W911NF-25-2-0040. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein. Kaleen Shrestha is supported by an NSF CISE Graduate Fellowship CSGrad4US under Grant No. 2313998 (Award ID G-2A-061).

References

Abdelnabi, S.; Gomaa, A.; Sivaprasad, S.; Schönherr, L.; and Fritz, M. 2024. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. *Advances in Neural Information Processing Systems*, 37: 83548–83599.

Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. Technical report, Anthropic.

Antonioni, D. 1998. Relationship between the big five personality factors and conflict management styles. *International journal of conflict management*, 9(4): 336–355.

Ashtiani, M. N.; and Raahemi, B. 2023. News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review. *Expert Systems with Applications*, 217: 119509.

Baumeister, R. F.; Vohs, K. D.; and Funder, D. C. 2007. Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on psychological science*, 2(4): 396–403.

Bell, E. C.; and Blakeney, R. N. 1977. Personality correlates of conflict resolution modes. *Human Relations*, 30(9): 849–857.

Brett, J. M.; Shapiro, D. L.; and Lytle, A. L. 1998. Breaking the bonds of reciprocity in negotiations. *Academy of Management Journal*, 41(4): 410–424.

Goldberg, L. R. 1992. The development of markers for the Big-Five factor structure. *Psychological assessment*, 4(1): 26.

Google. 2025. Gemini 2.0 Flash. Technical report, Google Cloud.

Graziano, W. G.; Jensen-Campbell, L. A.; and Hair, E. C. 1996. Perceiving interpersonal conflict and reacting to it: the case for agreeableness. *Journal of personality and social psychology*, 70(4): 820.

Gui, G.; and Toubia, O. 2023. The Challenge of Using LLMs to Simulate Human Behavior: A Causal Inference Perspective. *arXiv preprint arXiv:2312.15524*.

Hale, J. A.; Rakshit, S.; Chawla, K.; Brett, J. M.; and Gratch, J. 2025. Kodis: A multicultural dispute resolution dialogue corpus. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 12771–12785.

Han, B.; Kwon, D.; Lin, S.; Shrestha, K.; and Gratch, J. 2025. Can LLMs Generate Behaviors for Embodied Virtual Agents Based on Personality Traits? In *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, 1–10.

Huang, Y. J.; and Hadfi, R. 2024. How Personality Traits Influence Negotiation Outcomes? A Simulation based on Large Language Models. *arXiv preprint arXiv:2407.11549*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jiang, G.; Xu, M.; Zhu, S.-C.; Han, W.; Zhang, C.; and Zhu, Y. 2023. Evaluating and Inducing Personality in Pre-trained Language Models. arXiv:2206.07550.

Jones, R. E.; and Melcher, B. H. 1982. Personality and the preference for modes of conflict resolution. *Human relations*, 35(8): 649–658.

Kaushal, R.; and Kwantes, C. T. 2006. The role of culture and personality in choice of conflict management strategy. *International journal of intercultural relations*, 30(5): 579–603.

Kelley, H. H. 1996. *A classroom study of the dilemmas in interpersonal negotiations*. Berkeley Institute of International Studies.

Kwon, D.; Hae, J.; Clift, E.; Shamsoddini, D.; Gratch, J.; and Lucas, G. 2025. ASTRA: A Negotiation Agent with Adaptive and Strategic Reasoning via Tool-integrated Action for Dynamic Offer Optimization. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 16228–16249. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Li, Z.; Zhu, H.; Lu, Z.; Xiao, Z.; and Yin, M. 2025. From Text to Trust: Empowering AI-assisted Decision Making with Adaptive LLM-powered Analysis. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–18.

Lin, I. W.; Sharma, A.; Rytting, C. M.; Miner, A. S.; Suh, J.; and Althoff, T. 2024. IMBUE: improving interpersonal effectiveness through simulation and just-in-time feedback with human-language model interaction. *arXiv preprint arXiv:2402.12556*.

Mao, S.; Wang, X.; Wang, M.; Jiang, Y.; Xie, P.; Huang, F.; and Zhang, N. 2024. Editing Personality for Large Language Models. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 241–254. Springer.

Park, H.; and Antonioni, D. 2007. Personality, reciprocity, and strength of conflict resolution strategy. *Journal of research in personality*, 41(1): 110–125.

Park, J. S.; Popowski, L.; Cai, C.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User*

Interface Software and Technology, UIST '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393201.

Putnam, L. L.; and Jones, T. S. 1982. Reciprocity in negotiations: An analysis of bargaining interaction. *Communication monographs*, 49(3): 171–191.

Ross, L.; and Stillinger, C. 1991. Barriers to conflict resolution. *Negotiation journal*, 7(4): 389–404.

Serapio-García, G.; Safdari, M.; Crepy, C.; Sun, L.; Fitz, S.; Romero, P.; Abdulhai, M.; Faust, A.; and Matarić, M. 2023. Personality Traits in Large Language Models. arXiv:2307.00184.

Sorokovikova, A.; Fedorova, N.; Rezagholi, S.; and Yamshchikov, I. P. 2024. LLMs Simulate Big Five Personality Traits: Further Evidence. *arXiv preprint arXiv:2402.01765*.

Sternberg, R. J.; and Soriano, L. J. 1984. Styles of conflict resolution. *Journal of Personality and Social Psychology*, 47(1): 115.

Tehrani, H. D.; and Yamini, S. 2020. Personality traits and conflict resolution styles: A meta-analysis. *Personality and Individual Differences*, 157: 109794.

Ury, W. L.; Brett, J. M.; and Goldberg, S. B. 1988. *Getting disputes resolved: Designing systems to cut the costs of conflict.* Jossey-bass.

Wood, V. F.; and Bell, P. A. 2008. Predicting interpersonal conflict resolution styles from personality characteristics. *Personality and individual differences*, 45(2): 126–131.

Xie, C.; Chen, C.; Jia, F.; Ye, Z.; Lai, S.; Shu, K.; Gu, J.; Bibi, A.; Hu, Z.; Jurgens, D.; et al. 2024. Can Large Language Model Agents Simulate Human Trust Behavior? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zartman, I. W.; and Faure, G. O. 2005. The dynamics of escalation and negotiation. *Escalation and negotiation in international conflicts*, 3–20.

Zhou, X.; Zhu, H.; Mathur, L.; Zhang, R.; Yu, H.; Qi, Z.; Morency, L.-P.; Bisk, Y.; Fried, D.; Neubig, G.; et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.