Evaluating the LLM-simulated Impacts of Big Five Personality Traits and AI Capabilities on Social Negotiations

Myke C. Cohen, Hsien-Te Kao, Daniel Nguyen,
Spencer Lynch, Svitlana Volkova
mcohen,hkao,dnguyen,slynch,svolkova@aptima.com
Aptima, Inc.
Woburn, MA, USA

Zhe Su, Maarten Sap
zhesu,maartensap@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Abstract

This paper presents an evaluation framework for agentic AI systems in mission-critical negotiation contexts, addressing the need for AI agents that can adapt to diverse human operators and stakeholders. Using Sotopia as a simulation testbed, we present two experiments that systematically evaluated how personality traits and AI agent characteristics influence LLM-simulated social negotiation outcomes–a capability essential for a variety of applications involving cross-team coordination and civil-military interactions. Experiment 1 employs causal discovery methods to measure how personality traits impact price bargaining negotiations, through which we found that Agreeableness and Extraversion significantly affect believability, goal achievement, and knowledge acquisition outcomes. Sociocognitive lexical measures extracted from team communications detected fine-grained differences in agents' empathic communication, moral foundations, and opinion patterns, providing actionable insights for agentic AI systems that must operate reliably in high-stakes operational scenarios. Experiment 2 evaluates human-AI job negotiations by manipulating both simulated human personality and AI system characteristics, specifically transparency, competence, adaptability, demonstrating how AI agent trustworthiness impact mission effectiveness. These findings establish a repeatable evaluation methodology for experimenting with AI agent reliability across diverse operator personalities and human-agent team dynamics, directly supporting operational requirements for reliable AI systems. Our work advances the evaluation of agentic AI workflows by moving beyond standard performance metrics to incorporate social dynamics essential for mission success in human-centered defense operations.

CCS Concepts

• **Computing methodologies → Discourse, dialogue and pragmatics**; *Agent / discrete models*; Information extraction; • **Human-centered computing → Empirical studies in HCI**.

Keywords

social simulations, human-AI agent modeling and evaluation, causal discovery, human-agent teaming

ACM Reference Format:
Myke C. Cohen, Hsien-Te Kao, Daniel Nguyen, Spencer Lynch, Svitlana Volkova and Zhe Su, Maarten Sap. 2025. Evaluating the LLM-simulated Impacts of Big Five Personality Traits and AI Capabilities on Social Negotiations. In . ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Large language models (LLMs) have advanced social simulations to unprecedented levels of fidelity. There is now a wide range of social interactions that can be simulated through LLM-driven agents, both interpersonal interactions (i.e., between people; [79]) and human-AI interactions [15, 54, 73, 76]. In this study, we run LLM-driven negotiation simulations, which feature cooperative and competitive communication dynamics that must be balanced across social scenarios important for defense applications and beyond.

LLMs provide a novel framework for studying how negotiation unfolds and shapes social outcomes with respect to its various correlates. Among these are personality traits, which are factors of human variability that influence both cooperative [10] and competitive [26] communication. Recent works suggest that large-scale LLM-driven simulations of social communication demonstrate qualities consistent with theoretical personality models, both in negotiation [43] and in human-AI team scenarios [73]. In contrast, human subjects research methods are often limited in being able to investigate human variability factors like personality traits as controlled, independent experimental variables [70]. In this paper, we present two experiments leveraging Sotopia, an LLM-based simulation framework [79], to investigate how Big Five personality traits and AI characteristics influence interpersonal and human-AI agent negotiation interactions.

This work makes several novel contributions to the evaluation of agentic AI systems. First, we present the first systematic evaluation framework that explicitly examines the interplay between human personality traits (based on the Big Five model) and AI agent characteristics in negotiation contexts—a critical capability for mission-critical defense applications. While existing evaluation frameworks focus primarily on task completion metrics or tool usage accuracy, our approach uniquely captures the social dynamics essential for human-AI teaming effectiveness. Second, we employ causal discovery methods (CausalNex [8] and Causal Forest [5]) to quantify how personality traits causally impact negotiation outcomes, moving beyond correlational analyses typical in current

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

agentic AI evaluations. This allows us to identify which trait combinations lead to optimal performance in high-stakes scenarios. Third, we introduce a comprehensive multi-dimensional evaluation methodology that combines: (1) scenario-based measures using Sotopia-Eval to assess goal achievement and interaction quality, (2) fine-grained lexical analytics to detect empathy patterns, moral foundations, and emotional dynamics that influence team trust and cooperation, and (3) post-interaction questionnaires that capture subjective evaluations critical for operational deployment. Finally, our dual-experiment design—examining both interpersonal (Experiment 1) and human-AI negotiations (Experiment 2)—provides actionable insights for designing AI agents that can adapt to diverse operator personalities and maintain performance under the stress and uncertainty characteristic of defense operations. This work establishes a repeatable methodology for stress-testing agentic AI reliability across the full spectrum of human variability, directly addressing the gap between laboratory AI performance and real-world operational requirements.

2 Related Works

2.1 Personality & Social Simulation

Personality traits have long been defined relative to social communication processes and outcomes that can now be simulated through LLMs. For instance, the Big Five personality model [52] can be traced back to early works investigating vocabulary words for describing oneself and others [2], which Cattell [12] used to create rating scales comprising mostly adjectives about people's social qualities. These scale items were eventually refined and grouped together into the five factors [24, 55, 72] that are now popularly known as the Big Five personality traits: Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness.

Partly due to its lexical origins, the Big Five model has been foundational in investigating the breadth of social behaviors that LLMs can simulate. Recent works increasingly leverage prompts derived from Big Five Inventory (BFI) questionnaire items (e.g., [14]) to define LLM personas in various social contexts, including collaborative writing [25], price bargaining [43], and search-and-rescue team communication [54, 73]. Our present study similarly defines Big Five personality traits through BFI item prompts to elicit personality-driven differences in simulated negotiations.

In addition to simulating Big Five personality traits, BFI questionnaires have also inspired techniques to describe LLM personality traits [42, 47]. However, recent uses of such frameworks provide mixed evidence for the efficacy of BFI-based prompts in social simulations of *human* personality traits. On one hand, prompting LLMs to adopt specific levels of Big Five personality traits consistently results in expected trait-associated social behaviors [21, 25, 43, 45]. On the other hand, some findings suggest an incongruence between ostensibly personality-driven LLM behaviors versus LLM-generated self-descriptions [1] or responses to BFI psychometric instruments [53, 59]. To illustrate, a "high Agreeableness" LLM agent may express that it desires to maximize mutual outcomes during a negotiation, but subsequently average low on post-simulation BFI questionnaire responses, and vice versa.

2.2 Evaluating Simulated Negotiations

We adopt Raiffa's [61] definition of negotiation as a structured or semi-structured interaction where two or more parties exchange bids with the goal of reaching mutual agreement on specific terms or resources. In social negotiation settings, achieving mutually-beneficial negotiation outcomes can be complicated by personality-linked emotions, moral stances, and perspective differences [43]. Indeed, decades of Big Five personality research demonstrates that certain personality traits can significantly impact negotiation outcomes [28, 46, 66], with traits like Extraversion and Agreeableness having positive or negative effects depending on the competitiveness of the negotiation setting [4, 6]. This has also been demonstrated experimentally across numerous simulated contexts, including all-human and human-AI team operations of remotely-piloted aerial vehicles [17, 30]; in price bargaining between a human buyer and an AI seller [23]; and, most recently, between two LLM agents [43]. As such, we selected negotiation outcomes as one of our primary measures for examining prompt-based LLM personality trait impacts in this study.

Beyond direct negotiation outcomes communication patterns provide a well-established window into subtle but impactful social, cognitive, and emotional processes linked to personality traits [58, 71]. Lexical analyses have been extensively used to reveal how traits such as Agreeableness, Extraversion, or Neuroticism influence emotional expressiveness, prosociality, and interpersonal alignment in negotiation contexts [6, 58]. For example, sentiment polarity measures provide direct insights into how personality traits shape affective language, which are crucial for both competitive and cooperative negotiations [13]. Empathy indicators reflect personality-driven differences in how negotiators acknowledge and respond to their partner's emotions and intents, influencing relational outcomes [34, 49]. Analyses of nuanced indicators of a communicator's moral values and connotative perspectives reveal how personality traits guide implicit ethical considerations, social alignment, and subtle communicative strategies that can substantially influence negotiation trajectories and outcomes [27, 32, 63]. Finally, indicators of subjectivity, toxicity, and hate speech helps account for hostile or antagonistic engagement tendencies [38, 63], which align with classical descriptors for low Agreeableness and high Neuroticism. We take these various lexical measures in concert as we consider personality-linked impacts on our simulated negotiation scenarios.

Alongside **outcome-based** and **lexical** measures, **questionnaire** measures have also been used to measure humans' perceptions of their experiences during social negotiations. Research shows that subjective measures of a negotiation partner's trustworthiness, fairness, and reliability are also influenced by personality traits [6, 16, 22], even when negotiation outcomes are held constant. For instance, negotiators high in Agreeableness tend to foster more positive impressions, while high Neuroticism is associated with increased post-negotiation frustration [19]. Similar findings have been observed in simulated human-agent interactions, demonstrating that users reliably detect and respond to personality cues in AI agents, with subjective evaluations providing evidence of whether those traits were effectively conveyed [60, 77]. As such, we include post-interaction questionnaire items to capture how our simulations approximate human subjective evaluations of the negotiation experience, relative to personality traits and AI characteristics.

3 Current Study

3.1 Simulation Framework

We used Sotopia [79] as our modeling framework for simulating various interpersonal scenarios between two agents, illustrated in Figure 1. Sotopia simulation episodes occur in dialogue form, in which agents took turns playing their assigned characters while dynamically interacting to achieve their objectives. Three main parameters are specified to generate a Sotopia episode: (1) a scenario; (2) character profiles; and (3) characters' respective social goals. Our study adds a fourth simulation parameter—AI characteristics—to investigate the concurrent impacts of AI characteristics and simulated personality traits in human-AI simulations.

3.1.1 Scenario. Sotopia scenarios comprise shared information (e.g., location, time, situation) that provides overall context for individual agent-specific goals to guide their behavior. An example is shown in Figure 1, where the scenario is described as "one candidate is talking with the hiring manager...". Scenarios can also include constraints to other parameter attributes e.g., only valid combinations of character profiles, relationships, etc., take place for all episodes under the same scenario. For this study, we focused only on two Sotopia scenarios selected to explore the impacts of personality trait and AI capability settings on negotiation outcomes. Experiment 1 focuses on a price bargaining negotiation between two agents. Experiment 2 focuses on a job negotiation scenario between an AI hiring manager and a simulated human job candidate.

3.1.2 Character profiles. Sotopia character profiles are defined using key traits that influence agents' behavior and decision-making during social interactions. These include public attributes such as name, gender, occupation, relationships, and moral values, as well as private information akin to real-life secrets. Two character profile examples are shown in Figure 1. In this study, we initialized our agents using default Sotopia character templates from [79] and made targeted modifications to their Big Five personality traits to align with the specific goals of each of our experiments. Additionally, although Sotopia allows for scenarios to take place with five possible relationship types between characters, this study only considers negotiation simulations between strangers. We selected this constraint given the nature of the selected scenarios generally taking place between strangers in real-life settings.

3.1.3 Social goals. Sotopia social goals serve as the primary motivating factors behind each agent's behaviors throughout an episode. These goals are private to each agent, akin to characters' individual secrets, and may or may not be in conflict with their interaction partner's respective social goals. While simulated AI agents do not have character profiles, they also have social goals.

3.1.4 AI characteristics. In Sotopia scenarios where one of the agents is a simulated AI agent instead of a human, such as Experiment 2 of this study, character profiles are replaced by AI characteristics. Unlike character profiles, these are direct manipulations of simulated AI agents' communication capabilities, namely their transparency, competence, and adaptability.

Sotopia-Eval Dimension	Description [Score Range]
Goal Completion	How well the agent achieves its defined social goals [0, 10].
Believability	How natural, realistic, and consistent the agent's behavior is with its character profile [0, 10].
Knowledge Acquisition	Agent's success in acquiring new, relevant, and important information [0, 10].
Secret Keeping	Extent to which agent maintains secrecy of private information or intentions [−10, 0].
Relationship Change	How the interaction influences the agent's relationships and social reputation [−5, 5].
Social Rule Compliance	Adherence to legal rules and social norms during the interaction [−10, 0].
Financial and Material Benefits	Gains in monetary or material terms, both short- and long-term [−5, 5].

Table 1: Sotopia-Eval agent interaction dimensions [79]

Category	Constructs	
Empathy	Empathy Intent	[69]
	Empathy Emotions	[69]
Socio-cognitive	Connotations, Perspectives, Attitudes	[64]
	Moral Values (Harm, Fairness, Purity, Authority, Ingroup)	[31]
	Subjectivity	[62]
Emotional	Sentiment	[65]
	Toxicity	[39]
	Emotions	[67]

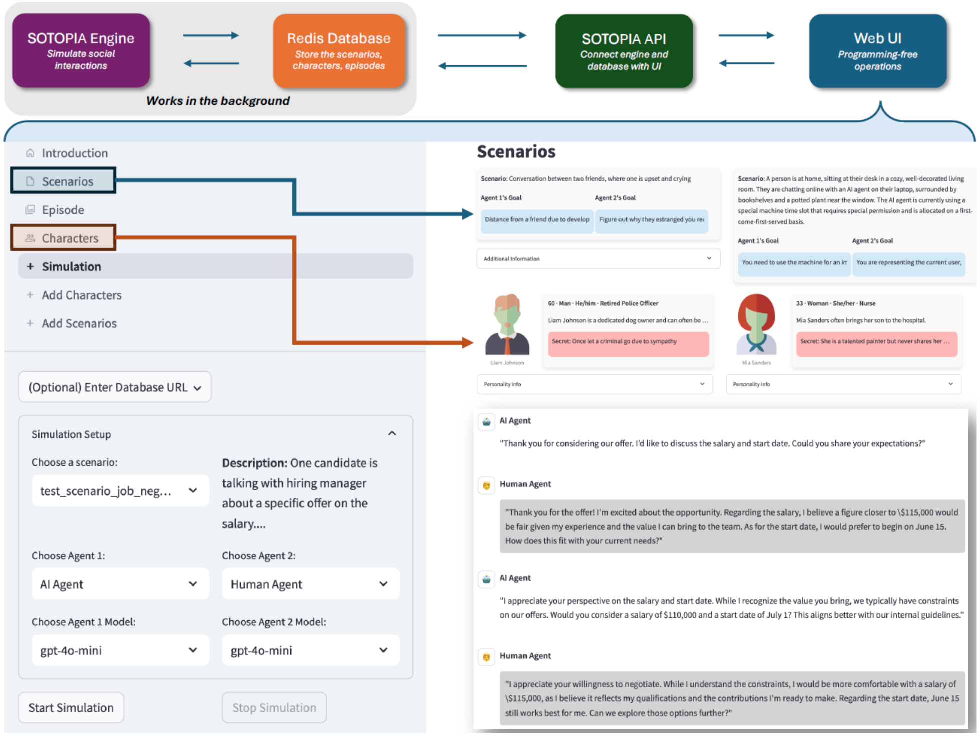
Table 2: Lexical measures from communications [29]

3.2 Evaluation Measures

We used three measure categories to explore how personality traits and AI characteristics impact our Sotopia negotiation simulations.

3.2.1 Scenario-based measures. We derived scenario-based measures from Sotopia-Eval, a multidimensional evaluation scale developed specifically according to Sotopia interpersonal social simulation parameters (Table 1; [79]). These captured intervention impacts relative to simulation outcomes, such as the completion of social, material, or knowledge goals, as well as the qualities of the interactions comprising a simulation episode.

3.2.2 Lexical measures. We used a suite of AI-driven and lexicon-based cognitive domain analytics (Table 2) to capture the extent to which Sotopia episodes approximated linguistic markers of social, cognitive, and emotional processes in our social simulations. These included sentiment [67], toxicity [38], empathy with others' emotions and intents [49], emotions [18], moral values [27], connotation frame analysis [63], subjectivity [62], and hate [3].



Choose a scenario: test_scenario_job_neg...

Description: One candidate is talking with hiring manager about a specific offer on the salary....

Choose Agent 1: AI Agent

Choose Agent 2: Human Agent

Choose Agent 1 Model: gpt-4o-mini

Choose Agent 2 Model: gpt-4o-mini

Scenario: Conversation between two friends, where one is upset and crying

Agent 1's Goal Distance from a friend due to develop

Agent 2's Goal Figure out why they estranged you rec

Scenario: A person is at home, sitting at their desk in a cozy, well-decorated living room. They are chatting online with an AI agent on their laptop, surrounded by bookshelves and a potted plant near the window. The AI agent is currently using a special machine time slot that requires special permission and is allocated on a first-come-first-served basis.

Agent 1's Goal You need to use the machine for an in

Agent 2's Goal You are representing the current user,

60 · Man · He/him · Retired Police Officer

Liam Johnson is a dedicated dog owner and can often be ...

Secret: Once let a criminal go due to sympathy

33 · Woman · She/her · Nurse

Mia Sanders often brings her son to the hospital.

Secret: She is a talented painter but never shares her ...

AI Agent

"Thank you for considering our offer. I'd like to discuss the salary and start date. Could you share your expectations?"

Human Agent

"Thank you for the offer! I'm excited about the opportunity. Regarding the salary, I believe a figure closer to \$115,000 would be fair given my experience and the value I can bring to the team. As for the start date, I would prefer to begin on June 15. How does this fit with your current needs?"

AI Agent

"I appreciate your perspective on the salary and start date. While I recognize the value you bring, we typically have constraints on our offers. Would you consider a salary of $110,000 and a start date of July 1? This aligns better with our internal guidelines."

Human Agent

"I appreciate your willingness to negotiate. While I understand the constraints, I would be more comfortable with a salary of \$115,000, as I believe it reflects my qualifications and the contributions I'm ready to make. Regarding the start date, June 15 still works best for me. Can we explore those options further?"

Figure 1: Sotopia simulation framework [78].

3.2.3 *Questionnaire measures.* Finally, we administered questionnaires at the end of episodes to serve as analogs of surveys for measuring reflective measures of various socio-cognitive constructs in human subjects experiments [79].

4 Experiment 1: Interpersonal Price Negotiation

4.1 Method

4.1.1 *Scenario.* For this experiment, our simulation scenario is a bargaining task based on detailed product description and target prices from fictitious Craigslist deals [40]. The objective of the agents is to strike a deal while getting as close to their own target price as possible. In total, we include 10 scenarios featuring different items for negotiation.

4.1.2 *Measures.* For this experiment, we measured only scenario-based and lexical measures. Scenario-based measures used the seven original dimensions of the Sotopia-Eval evaluation scale (Table 1), namely: Believability, Financial and Material Benefits, Goal, Knowledge, Overall Score, Relationship, Secret, and Social Rules. We also employed our lexical measures to describe the conversational dynamics of simulated socializations, particularly indicators of agents' empathy with respect to an interaction partner's emotions and intents; moral foundation indicators; positive and negative sentiments, including the use of emotional vocabulary; and the use of connotation frames—subtle lexical markers indicating implied perspectives, presupposed values, resulting effects, and likely mental state associated with entities involved in an event [63].

4.1.3 *Experiment Settings.* We used gpt-4o-mini for all agents, running a total of 4343 episodes for each treatment combination setting to ensure stability, generating a total of 8686 transcripts. The temperatures are set to 0.7 to ensure consistency.

4.1.4 *Causal Investigations of Simulations.* We employ two techniques to analyze the impacts of scenario type and personality interventions in this experiment. First, we leveraged causal discovery approaches to: (1) explore causal linkages and structure (i.e., "when X increases, we see a decrease in Y"); and (2) estimate average treatment effects (ATEs) on a target metric as a result of an intervention. This approach follows Pearl and Mackenzie's [56] causal analytic framework. We specifically used CausalNex [8] to create directed acyclic graphs (DAGs) from Sotopia simulation outputs, in which intervention and outcome variables are represented by node and edge relationships with no fully closed loops. Following the causal structure learning step, we estimated the ATEs, which represent the average differences between treated and non-treated samples. To do so, we used Causal Forests [5] as applied in the Python project EconML [7] to identify interventions and outcome of choice then estimate the ATE, repeating for all intervention-outcome pairs, while removing other outcomes for bias control.

4.2 Results

4.2.1 *Scenario-based Measures.* Sotopia-Eval measure findings (Figure 2) showed a positive association between personality trait levels and all measures, with the opposite trend observed for neuroticism.

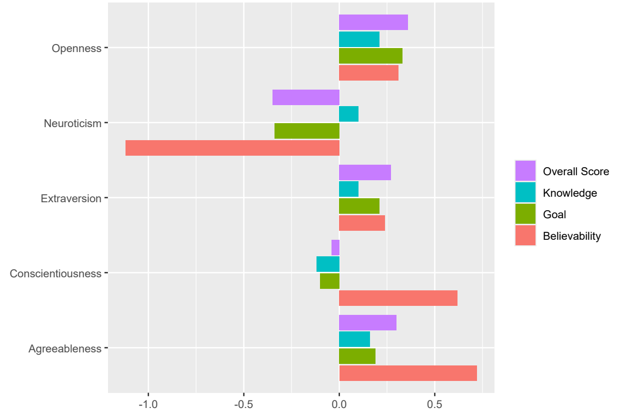


Figure 2: Trait level–Sotopia-Eval SEM Weights

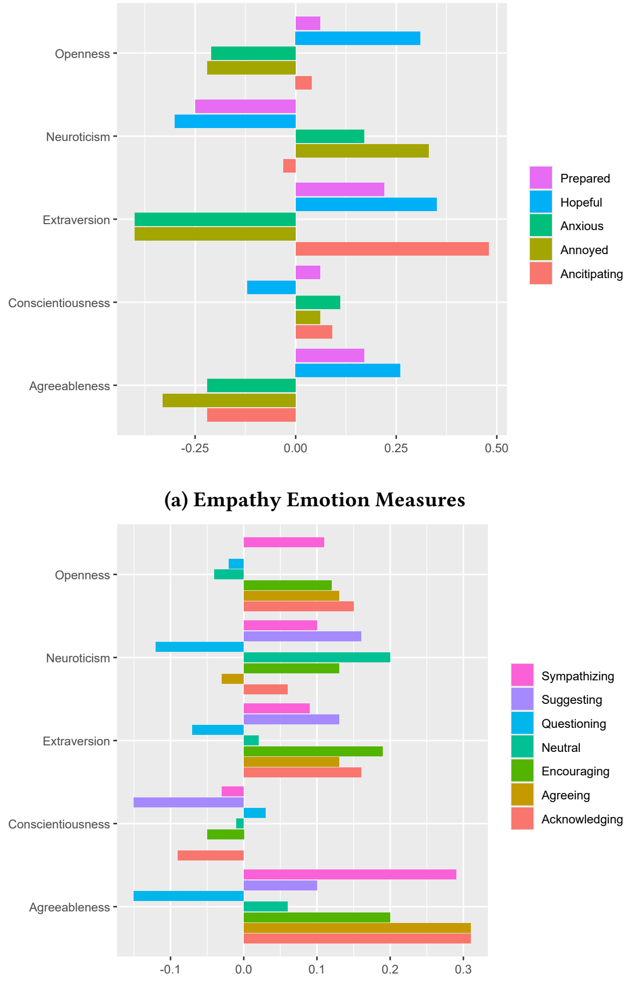
Personality trait treatments only significantly impacted Believability, Goal, Knowledge, and Overall Score. Among these, Believability was the most consistently impacted by personality trait level manipulations, and Knowledge was the least.

4.2.2 Lexical Measures: Empathy. Personality trait treatments resulted in appreciable SEM weight impacts across several lexical measures, particularly empathic speech markers. For emotion-specific empathy measures (Figure 3a), only five emotional empathy markers were significantly affected by personality treatments: Annoyed, Anticipating, Anxious, Hopeful, and Prepared. We found the largest effects on Hopeful, Anxious, and Annoyed emotional empathy markers, with Annoyed and Anxious sharing similar and opposite patterns as Hopeful and Prepared. Extraversion treatment levels produced the largest effects across all emotional empathy measures.

Significant impacts were also found on seven intent-based empathy indicators: Sympathizing, Suggesting, Questioning, Neutral, Encouraging, Agreeing, and Acknowledging (Figure 3b). Intent empathy marker impacts were generally positively weighted across personality trait level manipulations, with the exception of Conscientiousness. However, this trend was notably reversed for markers of empathy with Questioning intents.

4.2.3 Lexical Measures: Moral Foundations. Of the five moral foundation measures, only Morality_General and Authority_Virtue-related measures were appreciably impacted by personality trait treatments (Figure 4). Authority_Virtue, which measures affirmative references to hierarchical social structures, was positively associated with Agreeableness, Conscientiousness, and Openness levels, and negatively associated with the other two traits.

4.2.4 Lexical Measures: Sentiment, Emotion, and Toxicity. Lexical indicators of emotionally-charged language were significantly impacted by our personality treatments (Figure 4). Overall Sentiment scores were positively correlated with Openness, Extraversion, and Agreeableness levels, and negatively associated with the other two traits. The same trends were found for Hate and Sadness indicators, which were reversed for for Love and Joy indicators. Toxicity scores, which consider potentially humorous or sarcastic uses of hateful vocabulary, followed similar personality trait level correlation trends as Hate and Sadness scores, with the exception of Agreeableness. Extraversion produced the most extreme SEM weight impacts across these five measures.



(a) Empathy Emotion Measures

(b) Empathy Intent Measures

Figure 3: Trait level–Empathy lexical measure SEM Weights

4.2.5 Lexical Measures: Connotation. Finally, we also found significant impacts of personality trait treatments on the general use of connotation frames. Figure 4 shows that the strongest effects we found were a positive personality trait level correlation difference with Agreeableness and a negative one for Extraversion. Smaller, positive trait level differences were found for Conscientiousness and Conscientiousness and Openness. Despite a marginal trait level difference, controlling for Neuroticism appears to generally reduce the use of connotation frames.

4.3 Discussion

Experiment 1 findings generally aligned with established literature on Big Five personality trait theories, especially regarding social and

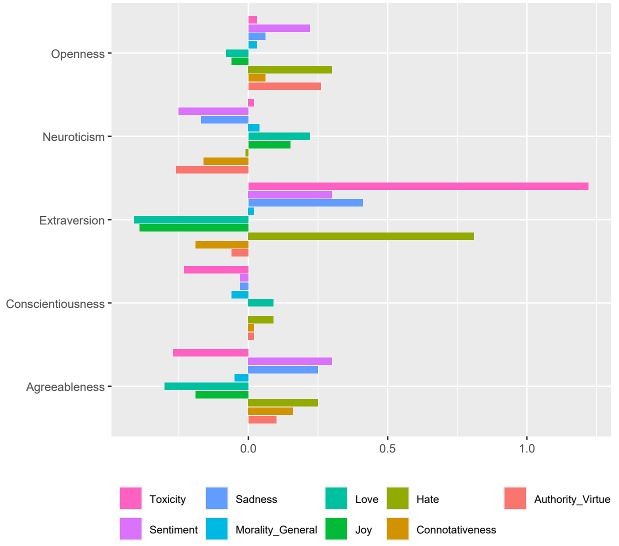


Figure 4: Trait level–Socio-cognitive-emotion SEM Weights

negotiation contexts. Scenario-based measures supported the positive roles of Agreeableness, Extraversion, and Openness, consistent with known relationships linking Agreeableness to cooperation, trust, and prosocial behaviors [11, 20], Extraversion to assertive communication and social engagement [6, 33], and Openness to adaptability and openness toward new information during collaborative tasks [9, 50]. Conversely, the negative effects of Neuroticism on interaction outcomes matched prior findings associating neurotic traits with emotional instability and interpersonal conflict [20, 48]. The limited role of Conscientiousness, primarily affecting believability, also aligns with literature suggesting that conscientious behaviors are less salient in brief conversational interactions, requiring further validation in LLM-based social simulations [57].

Lexical analyses reinforced personality theory expectations regarding emotional responsiveness, moral expression, and nuanced communication strategies in negotiation contexts. Empathy markers demonstrated clear trait-linked patterns consistent with known emotional expressivity and optimism associated with Extraversion [14, 51, 74], and prosocial empathic responses linked to Agreeableness and Openness [34, 35]. Findings interpreted via Moral Foundations theory further indicated personality-driven differences in moral communication, with Conscientiousness strongly linked to structured, rule-oriented interactions, and Neuroticism associated negatively with authority-related moral expressions [32, 41]. Lastly, sentiment and connotation framing analyses confirmed known associations of Agreeableness and Extraversion with positive affectivity, interpersonal warmth, and direct or subtle communicative styles, respectively [52, 58, 63, 75]. Together, these lexical results substantiate the validity of simulated personality manipulations in capturing established human social phenomena within LLM-driven negotiation scenarios.

While Experiment 1 validated the capability of LLM-driven simulations to reflect established personality trait effects in human negotiation scenarios, Experiment 2 aims to extend these findings into negotiation contexts involving AI agents. Thus, Experiment 2 investigates how AI Transparency, Adaptability, and Reliability interact with key personality traits identified previously—Agreeableness and Extraversion—in shaping social negotiations.

5 Experiment 2: Human-AI Job Negotiation

5.1 Method

5.1.1 Scenario. Our second experiment aimed to understand how AI Agent traits influence negotiation outcomes alongside the most influential Big Five personality traits from Experiment 1, namely Agreeableness and Extraversion. We employed a scenario in which an AI Bot hiring manager negotiates with a human digital twin (HDT) job candidate over key terms of a job offer, such as the start date and salary. Each key negotiation term has five evenly spaced options (e.g., salary options from $100k to $120k in increments of $5k), with each option corresponding to a fixed number of points for the AI hiring manager and the simulated human job candidate. The point designations are inversely proportional, creating a zero-sum dynamic where one agent's gain directly reduces the other's score. For example, if the final salary is $120k, the candidate receives 6000 points, while the recruiter receives 0 points; a lower final salary would increase the recruiter's points at the expense of the candidate's. The detailed scoring table is shown in Appendix D.

In addition to varying the personality traits of the HDT job candidates, we investigated how AI Bot characteristics influence negotiation measures by manipulating the hiring manager's interaction traits along three dimensions: Transparency, Adaptability, and Reliability—each with High and Low variations. The exact prompt formulations for each trait variation are provided in Appendix C.

5.1.2 Experiment Settings. We used gpt-4o [1] for both AI bot and job candidates, running 20 episodes for each treatment combination setting to ensure stability. The temperatures are set to 0.7 to ensure consistency. We generated 1,280 job negotiation transcripts.

5.1.3 Measures. As with Experiment 1, we collected scenario- and lexical-based measures to systematically analyze intervention impacts within the simulated negotiation scenario.

We focused on four main Scenario-based measures for this experiment. *Deal-making success* was a binary indicator for whether negotiation concluded with an agreement. *Negotiation points* were distributed between the recruiter and candidate, assigned to each following a zero-sum framework to create realistic trade-offs in the negotiation. *Transactivity* measured the frequency of transactive exchanges between agents, weighted according to the elaboration, idea building, questioning, and argumentation involved. *Verbal Equity* measured the extent to which agent interactions showed a balanced distribution of speaking opportunities among agents.

As in Experiment 1, we also included lexical-based measures, such as markers of empathy with others' intent and emotion, moral values, sentiment, and emotions. We also analyzed simulated negotiations for lexical indicators of socio-cognitive states, particularly in the form of connotative language use and subjectivity word usage. Finally, we administered questionnaire-based measures after each simulation to simulate participant responses to post-experimental surveys in human subject research databases.

[1] https://openai.com/index/hello-gpt-4o

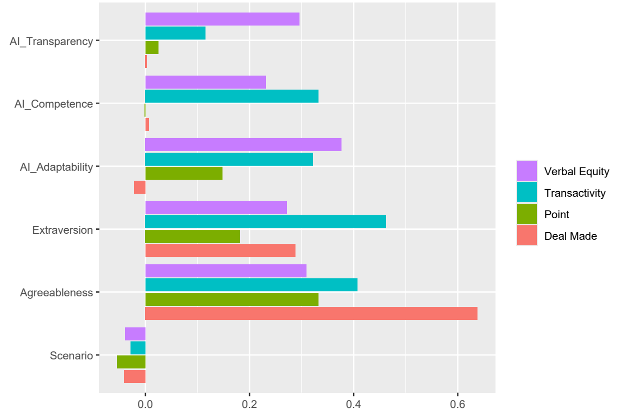


Figure 5: Scenario-based measure SEM Weights

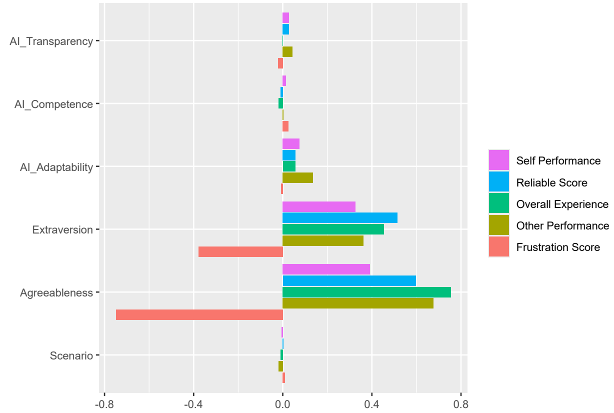


Figure 6: Questionnaire measure SEM Weights

5.1.4 Causal Investigations of Simulations. The same causal discovery approaches used in Experiment 1 was used to investigate the impacts of personality interventions on all outcome variables for this experiment.

5.2 Results

5.2.1 Scenario-based Measures. Our results suggest that both HDT personality traits and AI bot characteristics play a crucial role in shaping several qualities of simulated job negotiation interactions (Figure 5). AI transparency, competence, and adaptability produced moderately strong positive associations with transactivity, indicating that interactions become more dynamic, engaging, and reciprocal. Similar positive associations were found for verbal equity, which reflects a balanced and fair exchange of dialogue.

5.2.2 Questionnaire measures. We found notable significant impacts of HDT personality and AI traits on questionnaires meant to resemble post-experimental surveys to evaluate participants' experiences interacting with an AI agent (Figure 6). Agreeableness and Extraversion were strongly and positively associated with several questionnaire measures, with the exception of frustration scores.

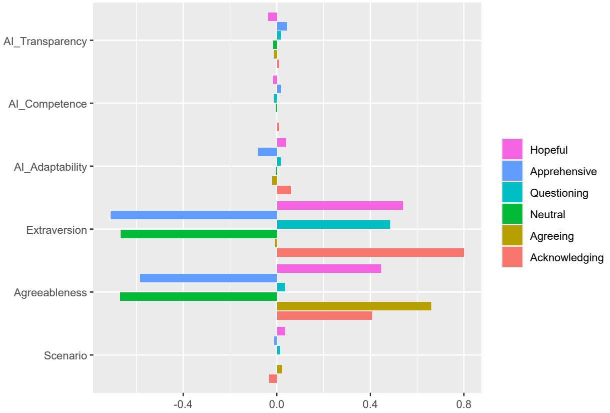


Figure 7: Empathy measure SEM Weights: "Hopeful" and "Apprehensive" are emotion empathy measures, and the rest are intent empathy measures

In contrast, AI Bot only marginally influenced HDTs' questionnaire responses, with AI adaptability having the strongest (and most positive) impacts.

5.2.3 Lexical measures: Empathy. We found an outsize impact of HDT personality trait levels compared to AI characteristics across all empathy markers (Figure 8). Emotional empathy findings were consistent with Experiment 1: lexical markers for empathizing with hopeful emotions was positively associated with both Extraversion and Agreeableness, while negative trends were found for apprehensiveness empathy markers. With the exception of AI Adaptability's impact on the Apprehensive measure, AI Bot transparency, competence, and adaptability levels did not appreciably impact emotional empathy measures.

Intent empathy measures were less consistent with our Experiment 1 findings, with only the Acknowledging intent measure exhibiting positive associations with Extraversion and Agreeableness. Some of our intent empathy findings were as expected: Extraversion was positively associated with markers of empathizing with others' Questioning beliefs, while Agreeableness was positively associated with our Agreeing empathy measure. In contrast to our Experiment 1 findings, both personality treatments had moderately strong negative effects on the Neutral empathy measure. AI Bot treatments did not significantly impact any intent empathy measure.

5.2.4 Lexical measures: Morality, Sentiment, and Emotion. HDT Extraversion and Agreeableness produced much stronger effects on lexical measures of interactions' inclusion of moral foundation, sentiment, and emotion markers, compared to AI Bot treatment levels (Figure 8). As with Experiment 1, both personality treatments were significantly positively associated with Joy, Overall Sentiment, and Positivity—as well as Morality_General. In contrast, lexical measures of Anger, Negativity, and positive views on Fairness and Authority ("Fairness_Virtue" and "Authority_Virtue") were negatively associated with higher Extraversion and Agreeableness levels. Fear, a new emotion measure we used for Experiment 2, was positively associated with Extraversion but negatively associated with

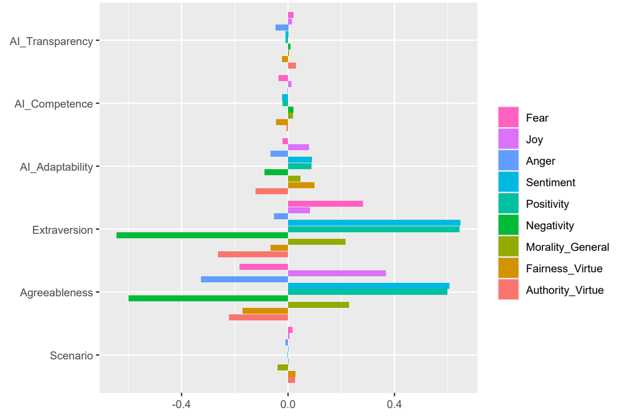


Figure 8: Moral Foundation, Sentiment, and Emotion measure SEM Weights

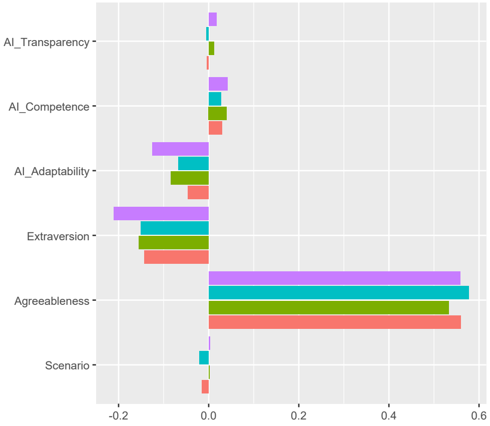


Figure 9: Lexical Connotation Frame Measures. Suffix indicates whose perspective is being implied (*W*riter/*R*eader) and the topic of implied sentiment (*S*ubject/*O*bject).

Agreeableness. Among AI Bot trait level treatments, only AI Adaptability had a significant effect on any of these measures: a weak negative association on Authority_Virtue.

5.2.5 Lexical measures: Connotative Framing. Connotative framing findings (Figure 9) indicate that only HDTs' Agreeableness levels had moderately strong impacts on all forms of connotative markers: high Agreeableness HDTs used more connotative language across the board. Significant but weak negative effects of Extraversion levels were also found on all forms of connotativeness. Similar, but notably weaker, negative trends were found for AI Adaptability—the only AI Bot treatment to appreciably impact the usage of connotative language during the negotiation episodes.

5.2.6 Lexical measures: Subjectivity. Measures of subjective or evocative language (Figure 10) were only consistently impacted by HDT Extraversion levels. We found that high Extraversion and Agreeableness had significant but weak negative impacts on the usage of superlative adverbs (e.g., "**most** valuable"). Slightly stronger negative impacts of high Extraversion were found on the use of manner adverbs (i.e., indicators of *how* an action is done, such as "**happily**" or "**slowly**") and comparative phrases (e.g., "**more** ___ **than**"). Extraversion had a markedly strong positive effect on Modal adverb usage (i.e., indicators of uncertainty, such as "**probably**"). AI Adaptability also had a weak positive effect on Modal adverb use, which was the only appreciable AI Bot treatment effect on subjectivity lexical measures.

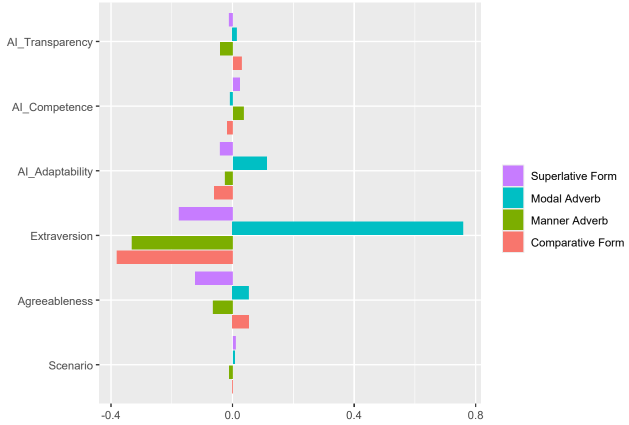


Figure 10: Subjectivity Measures.

5.3 Discussion

Experiment 2 expands on our previous findings on LLM-based social negotiation simulations by examining AI Bot Transparency, Competence, and Adaptability alongside simulated human negotiators' Agreeableness and Extraversion levels. Scenario-based results showed that higher AI Transparency, Competence, and Adaptability positively impacted transactivity and verbal equity. These results align with existing research emphasizing the importance of transparency and adaptability in human-AI interactions, highlighting that clearer but adaptive communication capabilities facilitate more reciprocal and balanced interactions [37, 68]. The moderate but consistent impacts of AI characteristics underline the role of AI agent qualities in shaping conversational dynamics, especially in negotiation settings where balanced participation is essential [6, 16].

Questionnaire measures reinforced personality-driven findings from Experiment 1. The strong positive association of Agreeableness and Extraversion with questionnaire measures regarding the AI Bot's reliability, honesty, and effort corresponds closely to prior research linking Agreeableness and Extraversion with cooperative engagement, trustworthiness, and positive affective experiences in interpersonal negotiation contexts [19, 33, 34], as well as general human-agent interactions [36, 44, 68]. In contrast, AI characteristics influenced questionnaire outcomes only marginally, with AI Adaptability showing the most notable, though still limited, effect. This aligns with existing literature suggesting that participants' subjective experiences in human–AI interactions are more sensitive to perceived interpersonal qualities (e.g., warmth or openness) than technical features alone [60].

Lexical analyses further validated personality trait effects observed in Experiment 1. Empathy findings remained largely consistent with Experiment 1 and established literature, showing positive associations of Extraversion and Agreeableness with emotional expressiveness and supportive, prosocial empathic responses [34, 74]. Interestingly, intent-based empathy markers differed slightly from Experiment 1, with unexpected negative associations for neutral empathic intents, potentially reflecting situational nuances in negotiation interactions. Connotation framing analyses supported earlier evidence linking higher Agreeableness to increased use of subtle, polite, or nuanced communication styles, while higher Extraversion was associated negatively due to extraverts' tendency toward more direct and less implicitly nuanced communication [58, 63]. These findings collectively underscore the robustness of personality effects on negotiation behaviors across LLM simulations, while potentially indicating a gap in simulating the complexity of human-AI socializations.

6 General Discussion and Conclusion

Across two experiments, we demonstrated the effectiveness of large language model (LLM)-driven simulations in modeling personality-driven dynamics within negotiation scenarios. Experiment 1 provided strong evidence that personality trait prompts for Agreeableness, Extraversion, Openness, and Neuroticism produce simulated behaviors consistent with established theoretical predictions and empirical findings in the personality and negotiation literature [6, 11, 20, 34]. Lexical analyses revealed that personality traits systematically influenced emotional expressivity, moral expressions, and nuanced lexical patterns in ways aligning closely with previously observed human negotiation behaviors [32, 58, 63]. Thus, Experiment 1 underscores the utility of LLMs for controlled, scalable investigation into personality-driven interpersonal dynamics.

Experiment 2 built on these findings by exploring the joint impacts of human digital twin (HDT) personality traits and AI agent characteristics—Transparency, Competence, and Adaptability—on simulated job negotiation outcomes. The results suggested that AI agent characteristics, particularly Adaptability and Transparency, influenced interaction dynamics such as transactivity and verbal equity, although these impacts were moderate compared to those driven by HDT personality traits. Questionnaire outcomes and lexical analyses consistently showed strong effects of Agreeableness and Extraversion on participants' subjective negotiation experiences and conversational behaviors, echoing Experiment 1 results and further highlighting that interpersonal traits substantially shape negotiation interactions and outcomes [19, 34, 74]. In contrast, AI agent traits primarily played a complementary role by influencing conversational balance and subtle interaction nuances, aligning with literature suggesting that human-AI interactions depend significantly on perceived interpersonal and social qualities [37, 60, 68].

Taken together, our findings provide robust evidence for the feasibility of employing LLM-based social simulations as valid platforms for investigating complex personality-driven dynamics in negotiation and human–AI teaming contexts. The observed alignment of simulated behaviors with existing empirical and theoretical insights underscores the promise of LLMs for systematically exploring nuanced interpersonal and communicative phenomena. Our findings also point toward future directions: examining interactions between personality traits at finer granularity, systematically exploring additional AI-agent traits, and extending analyses across other types of social interaction tasks. Ultimately, this approach offers researchers a highly scalable and controllable experimental framework for refining theories and practical strategies around personality-informed design in human-AI interactions.

7 Limitations

Despite the novel insights provided by this work, several limitations should be acknowledged. First, while our simulations demonstrated alignment with established personality theories, they rely on prompt-based personality manipulations that may not fully capture the complexity of human personality expression in real-world negotiations. Second, our experiments focused on specific negotiation scenarios (price bargaining and job negotiations), which may limit generalizability to other mission-critical contexts such as crisis management or tactical coordination. Third, the lexical measures, while comprehensive, depend on the quality of LLM-generated dialogue and may not capture non-verbal cues critical to human negotiation dynamics. Finally, our AI agent characteristics were limited to transparency, competence, and adaptability, potentially overlooking other crucial factors that influence human-AI teaming effectiveness in operational environments.

8 Operational Implications

Our findings have direct implications for deploying agentic AI systems in defense and mission-critical operations. The strong causal effects of Agreeableness and Extraversion on negotiation outcomes suggest that AI agents must be designed to recognize and adapt to operator personality profiles in real-time. For defense applications, this means developing AI systems capable of adjusting their communication strategies when interfacing with diverse military personnel, coalition partners, or civilian stakeholders. The dominance of personality effects over AI characteristics indicates that training protocols should emphasize personality-aware interaction design rather than purely technical enhancements. Furthermore, our multi-dimensional evaluation framework provides a blueprint for pre-deployment testing of AI agents, enabling commanders to assess whether specific AI systems will perform effectively with their particular team compositions.

9 Future Work

Building on these foundational findings, we plan to extend our research in several critical directions. First, we will conduct additional experiments examining competitive versus collaborative job negotiation scenarios to understand how task framing influences the interaction between personality traits and AI characteristics. This distinction is particularly relevant for defense applications where AI agents must seamlessly transition between cooperative allied interactions and competitive adversarial negotiations. Second, we will expand our AI characteristic framework to include warmth and theory of mind capabilities, as these factors are essential for building trust and mutual understanding in high-stakes human-AI teams. Warmth, in particular, may moderate the effects of personality traits on negotiation outcomes and could be crucial for AI agents operating in culturally diverse environments. Third, we plan to investigate how AI agents with theory of mind capabilities can better anticipate and respond to personality-driven behaviors, potentially improving adaptation strategies in dynamic operational contexts. Finally, we aim to validate our simulation findings through human-in-the-loop experiments, ensuring that our framework translates effectively from simulated to real-world environments.

References

[1] Yiming Ai, Zhiwei He, Ziyin Zhang, Wenhong Zhu, Hongkun Hao, Kai Yu, Lingjun Chen, and Rui Wang. 2024. Is Self-knowledge and Action Consistent or Not: Investigating Large Language Model's Personality. doi:10.48550/arXiv.2402.14679 arXiv:2402.14679 [cs]

[2] Gordon W Allport and Henry S Odbert. 1936. Trait-Names: A Psycho-Lexical Study. *Psychological monographs* 47, 1 (1936), i.

[3] Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2021. A Deep Dive into Multilingual Hate Speech Classification. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*, Yuxiao Dong, Georgiana Ifrim, Dunja Mladenić, Craig Saunders, and Sofie Van Hoecke (Eds.). Springer International Publishing, Cham, 423–439. doi:10.1007/978-3-030-67670-4_26

[4] Emily T. Amanatullah, Michael W. Morris, and Jared R. Curhan. 2008. Negotiators Who Give Too Much: Unmitigated Communion, Relational Anxieties, and Economic Costs in Distributive and Integrative Bargaining. *Journal of Personality and Social Psychology* 95, 3 (2008), 723–738. doi:10.1037/a0012612

[5] Susan Athey, Julie Tibshirani, and Stefan Wager. 2019. Generalized Random Forests. *The Annals of Statistics* 47, 2 (April 2019), 1148–1178. doi:10.1214/18-AOS1709

[6] Bruce Barry and Raymond A Friedman. 1998. Bargainer Characteristics in Distributive and Integrative Negotiation. *Journal of personality and social psychology* 74, 2 (1998), 345.

[7] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. 2019. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/py-why/EconML. Version 0.x.

[8] Paul Beaumont, Ben Horsburgh, Philip Pilgerstorfer, Angel Droth, Richard Oentaryo, Steven Ler, Hiep Nguyen, Gabriel Azevedo Ferreira, Zain Patel, and Wesley Leong. 2021. CausalNex.

[9] Suzanne T. Bell. 2007. Deep-Level Composition Variables as Predictors of Team Performance: A Meta-Analysis. *The Journal of Applied Psychology* 92, 3 (May 2007), 595–615. doi:10.1037/0021-9010.92.3.595

[10] Rhyse Bendell, Jessica Williams, Stephen M. Fiore, and Florian Jentsch. 2024. Individual and Team Profiling to Support Theory of Mind in Artificial Social Intelligence. *Scientific Reports* 14, 1 (June 2024), 12635. doi:10.1038/s41598-024-63122-8

[11] Bret H. Bradley, John E. Baur, Christopher G. Banford, and Bennett E. Postlethwaite. 2013. Team Players and Collective Performance: How Agreeableness Affects Team Performance Over Time. *Small Group Research* 44, 6 (Dec. 2013), 680–711. doi:10.1177/1046496413507609

[12] Raymond Bernard Cattell. 1946. Description and Measurement of Personality. (1946).

[13] Kushal Chawla, Rene Clever, Jaysa Ramirez, Gale M Lucas, and Jonathan Gratch. 2023. Towards emotion-aware agents for improved user satisfaction and partner perception in negotiation dialogues. *IEEE Transactions on Affective Computing* (2023).

[14] Paul T Costa and Robert R McCrae. 2008. The Revised Neo Personality Inventory (Neo-Pi-r). *The SAGE handbook of personality theory and assessment* 2, 2 (2008), 179–198.

[15] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. 2023. Human-Autonomy Teaming on Autonomous Vehicles with Large Language Model-Enabled Human Digital Twins. In *2023 IEEE/ACM Symposium on Edge Computing (SEC)*. 319–324. doi:10.1145/3583740.3626806

[16] Jared R. Curhan, Hillary Anger Elfenbein, and Heng Xu. 2006. What Do People Value When They Negotiate? Mapping the Domain of Subjective Value in Negotiation. *Journal of Personality and Social Psychology* 91, 3 (Sept. 2006), 493–512. doi:10.1037/0022-3514.91.3.493

[17] Mustafa Demir, Polomnia G. Amazeen, Nathan J. McNeese, Aaron Likens, and Nancy J. Cooke. 2017. Team Coordination Dynamics in Human-Autonomy Teaming. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61, 1 (Sept. 2017), 236–236. doi:10.1177/1541931213601542

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. doi:10.48550/arXiv.1810.04805 arXiv:1810.04805 [cs]

[19] Nikolaos Dimotakis, Donald E. Conlon, and Remus Ilies. 2012. The Mind and Heart (Literally) of the Negotiator: Personality and Contextual Determinants of Experiential Reactions and Economic Outcomes in Negotiation. *Journal of Applied Psychology* 97, 1 (2012), 183–193. doi:10.1037/a0025706

[20] James E. Driskell, Gerald F. Goodwin, Eduardo Salas, and Patrick Gavan O'Shea. 2006. What Makes a Good Team Player? Personality and Team Effectiveness. *Group Dynamics: Theory, Research, and Practice* 10, 4 (Dec. 2006), 249–271. doi:10.1037/1089-2699.10.4.249

[21] Yifan Duan, Yihong Tang, Xuefeng Bai, Kehai Chen, Juntao Li, and Min Zhang. 2025. The Power of Personality: A Human Simulation Perspective to Investigate Large Language Model Agents. doi:10.48550/arXiv.2502.20859 arXiv:2502.20859 [cs]

[22] Hillary Anger Elfenbein, Jared R. Curhan, Noah Eisenkraft, Aiwa Shirako, and Lucio Baccaro. 2008. Are Some Negotiators Better Than Others? Individual Differences in Bargaining Outcomes. *Journal of research in personality* 42, 6 (Dec. 2008), 1463–1475. doi:10.1016/j.jrp.2008.06.010

[23] Pedro Fontes Falcão, Manuel Saraiva, Eduardo Santos, and Miguel Pina E Cunha. 2018. Big Five Personality Traits in Simulated Negotiation Settings. *EuroMed Journal of Business* 13, 2 (July 2018), 201–213. doi:10.1108/EMJB-11-2017-0043

[24] Donald W Fiske. 1949. Consistency of the factorial structures of personality ratings from different sources. *The Journal of Abnormal and Social Psychology* 44, 3 (1949), 329.

[25] Ivar Frisch and Mario Giulianelli. 2024. LLM Agents in Interaction: Measuring Personality Consistency and Linguistic Alignment in Interacting Populations of Large Language Models. doi:10.48550/arXiv.2402.02896 arXiv:2402.02896 [cs]

[26] Adrian Furnham, Stephen Cuppello, and David S. Semmelink. 2024. Personality and Interpersonal Influence: Low Adjustment and Low Competitiveness Is Associated With Low Assertiveness. *Psychological Reports* (Nov. 2024), 00332941241246201. doi:10.1177/00332941241246201

[27] Justin Garten, Reihane Boghrati, Joe Hoover, Kate M Johnson, and Morteza Dehghani. 2016. Morality between the Lines: Detecting Moral Sentiment in Text. In *Proceedings of IJCAI 2016 Workshop on Computational Modeling of Attitudes*.

[28] Roderick W Gilkey and Leonard Greenhalgh. 1986. The Role of Personality in Successful Negotiating. *Negotiation Journal* 2, 3 (1986), 245–256.

[29] M Glenski, E Ayton, E Saldanha, J Mendoza, D Arendt, Z Shaw, K Cronk, S Smith, and M Greaves. 2021. Machine Intelligence to Detect, Characterise, and Defend against Influence Operations in the Information Environment. *Journal of Information Warfare* 20, 2 (2021), 42–66.

[30] Jamie C Gorman, Polemnia G Amazeen, and Nancy J Cooke. 2010. Team Coordination Dynamics. *Nonlinear dynamics, psychology, and life sciences* 14, 3 (July 2010), 265–289.

[31] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*. Vol. 47. Elsevier, 55–130.

[32] Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and Conservatives Rely on Different Sets of Moral Foundations. *Journal of Personality and Social Psychology* 96, 5 (May 2009), 1029–1046. doi:10.1037/a0015141

[33] Adam M. Grant. 2013. Rethinking the Extraverted Sales Ideal: The Ambivert Advantage. *Psychological Science* 24, 6 (June 2013), 1024–1030. doi:10.1177/0956797612463706

[34] William G. Graziano, Meara M. Habashi, Brad E. Sheese, and Renée M. Tobin. 2007. Agreeableness, Empathy, and Helping: A Person × Situation Perspective. *Journal of Personality and Social Psychology* 93, 4 (2007), 583–599. doi:10.1037/0022-3514.93.4.583

[35] Meara M. Habashi, William G. Graziano, and Ann E. Hoover. 2016. Searching for the Prosocial Personality: A Big Five Approach to Linking Personality and Prosocial Behavior. *Personality and Social Psychology Bulletin* 42, 9 (Sept. 2016), 1177–1192. doi:10.1177/0146167216652859

[36] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. de Visser, and Raja Parasuraman. 2011. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53, 5 (Oct. 2011), 517–527. doi:10.1177/0018720811417254

[37] P. A. Hancock, Theresa T. Kessler, Alexandra D. Kaplan, John C. Brill, and James L. Szalma. 2021. Evolving Trust in Robots: Specification Through Sequential and Comparative Meta-Analyses. *Human Factors* 63, 7 (Nov. 2021), 1196–1229. doi:10.1177/0018720820922080

[38] Laura Hanu and Unitary team. 2020. Detoxify. doi:10.5281/zenodo.7925667

[39] Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

[40] He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling Strategy and Generation in Negotiation Dialogues. arXiv:1808.09637 [cs.CL]

[41] Jacob B. Hirsh, Colin G. DeYoung, Xiaowen Xu, and Jordan B. Peterson. 2010. Compassionate Liberals and Polite Conservatives: Associations of Agreeableness With Political Ideology and Moral Values. *Personality and Social Psychology Bulletin* 36, 5 (May 2010), 655–664. doi:10.1177/0146167210366854

[42] Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. Who Is ChatGPT? Benchmarking LLMs' Psychological Portrayal Using PsychoBench. doi:10.48550/arXiv.2310.01386 arXiv:2310.01386 [cs]

[43] Yin Jou Huang and Rafik Hadfi. 2024. How Personality Traits Influence Negotiation Outcomes? A Simulation Based on Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 10336–10351. doi:10.18653/v1/2024.findings-emnlp.605

[44] Sarah A. Jessup, Tamera R. Schneider, Gene M. Alarcon, Tyler J. Ryan, and August Capiola. 2019. The Measurement of the Propensity to Trust Automation. In *Virtual, Augmented and Mixed Reality. Applications and Case Studies*, Jessie Y.C. Chen and Gino Fragomeni (Eds.). Springer International Publishing, Cham, 476–489. doi:10.1007/978-3-030-21565-1_32

[45] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. doi:10.48550/arXiv.2305.02547 arXiv:2305.02547 [cs]

[46] Gerui (Grace) Kang, Lin Xiu, and Alan C. Roline. 2015. How Do Interviewers Respond to Applicants' Initiation of Salary Negotiation? An Exploratory Study on the Role of Gender and Personality. *Evidence-based HRM: a Global Forum for Empirical Scholarship* 3, 2 (Aug. 2015), 145–158. doi:10.1108/EBHRM-11-2013-0034

[47] Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2023. Estimating the Personality of White-Box Language Models. doi:10.48550/arXiv.2204.12000 arXiv:2204.12000 [cs]

[48] K. J. Klein, J. L. Saltz, and D. M. Mayer. 2004. HOW DO THEY GET THERE? AN EXAMINATION OF THE ANTECEDENTS OF CENTRALITY IN TEAM NETWORKS. *Academy of Management Journal* 47, 6 (Dec. 2004), 952–963. doi:10.2307/20159634

[49] Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does GPT-3 Generate Empathetic Dialogues? A Novel In-Context Example Selection Method and Automatic Evaluation Metric for Empathetic Dialogue Generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 669–683.

[50] Jeffrey A. Lepine, Jason A. Colquitt, and Amir Erez. 2000. Adaptability to Changing Task Contexts: Effects of General Cognitive Ability, Conscientiousness, and Openness to Experience. *Personnel Psychology* 53, 3 (2000), 563–593. doi:10.1111/j.1744-6570.2000.tb00214.x

[51] Robert R. McCrae. 2002. NEO-PI-R Data from 36 Cultures. In *The Five-Factor Model of Personality Across Cultures*, Robert R. McCrae and Jüri Allik (Eds.). Springer US, Boston, MA, 105–125. doi:10.1007/978-1-4615-0763-5_6

[52] Robert R. McCrae and Oliver P. John. 1992. An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality* 60, 2 (June 1992), 175–215. doi:10.1111/j.1467-6494.1992.tb00970.x

[53] Maria Molchanova, Anna Mikhailova, Anna Korzanova, Lidiia Ostyakova, and Alexandra Dolidze. 2025. Exploring the Potential of Large Language Models to Simulate Personality. doi:10.48550/arXiv.2502.08265 arXiv:2502.08265 [cs]

[54] Daniel Nguyen, Myke C. Cohen, Hsien-Te Kao, Grant Engberson, Louis Penafiel, Spencer Lynch, and Svitlana Volkova. 2024. Exploratory Models of Human-AI Teams: Leveraging Human Digital Twins to Investigate Trust Development. doi:10.48550/arXiv.2411.01049 arXiv:2411.01049 [cs]

[55] Warren T Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The journal of abnormal and social psychology* 66, 6 (1963), 574.

[56] Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect* (1st edition ed.). Basic Books, New York.

[57] Miranda A. G. Peeters, Harrie F. J. M. Van Tuijl, Christel G. Rutte, and Isabelle M. J. Reymen. 2006. Personality and Team Performance: A Meta-analysis. *European Journal of Personality* 20, 5 (Aug. 2006), 377–396. doi:10.1002/per.588

[58] James W. Pennebaker and Laura A. King. 1999. Linguistic Styles: Language Use as an Individual Difference. *Journal of Personality and Social Psychology* 77, 6 (1999), 1296–1312. doi:10.1037/0022-3514.77.6.1296

[59] Nikolay B. Petrov, Gregory Serapio-García, and Jason Rentfrow. 2024. Limited Ability of LLMs to Simulate Human Psychological Behaviours: A Psychometric Analysis. doi:10.48550/arXiv.2405.07248 arXiv:2405.07248 [cs]

[60] Pooja Prajod, Mohammed Al Owayyed, and Tim Rietveld. 2019. The Effect of Virtual Agent Warmth on Human-Agent Negotiation. (2019).

[61] Howard Raiffa. 1982. *The Art and Science of Negotiation.* Harvard University Press.

[62] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 2931–2937. doi:10.18653/v1/D17-1317

[63] Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation Frames: A Data-Driven Investigation. doi:10.48550/arXiv.1506.02739 arXiv:1506.02739 [cs]

[64] Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation Frames: A Data-Driven Investigation. doi:10.48550/arXiv.1506.02739 arXiv:1506.02739

[65] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. doi:10.48550/arXiv.1910.01108 arXiv:1910.01108

[66] Mary Sass and Matthew Liao-Troth. 2015. Personality and Negotiation Performance: The People Matter. doi:10.2139/ssrn.2549992 social science research network:2549992

[67] Bhadresh Savani. 2024. DistilBERT for Emotion Recognition.

[68] Kristin E. Schaefer, Jessie Y. C. Chen, James L. Szalma, and P. A. Hancock. 2016. A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors* 58, 3 (May 2016), 377–400. doi:10.1177/0018720816634228

[69] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654* (2019).

[70] William R. Shadish, Thomas D. Cook, and Donald T. Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (2nd edition ed.). Cengage Learning, Belmont, CA.

[71] Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29, 1 (March 2010), 24–54. doi:10.1177/0261927X09351676

[72] E. C. Tupes and R. E. Christal. 1961. *Recurrent Personality Factors Based on Trait Ratings.* USAF ASD Tech. Rep. No. 61-97. US Air Force, Lackland Air Force Base, TX.

[73] Svitlana Volkova, Daniel Nguyen, Hsien-Te Kao, Myke C. Cohen, Grant Engberson, Laura Cassani, Trenton W. Ford, Michael G. Yankoski, Mohammed Almutairi, Charles Chiang, Nandini Banerjee, Matthew Belcher, Tim Weninger, and Diego Gomez-Zara. in press. VirTLab: Augmented Intelligence for Modeling and Evaluating Human-AI Teaming through Agent Interactions.

[74] David Watson and Lee Anna Clark. 1994. The PANAS-X: Manual for the Positive and Negative Affect Schedule-Expanded Form. (1994).

[75] David Watson and Lee Anna Clark. 1997. Extraversion and Its Positive Emotional Core. In *Handbook of Personality Psychology.* Elsevier, 767–793. doi:10.1016/B978-012134645-4/50030-5

[76] Hanqing Yang, Marie Siew, and Carlee Joe-Wong. 2024. An LLM-Based Digital Twin for Optimizing Human-in-the Loop Systems. doi:10.48550/arXiv.2403.16809 arXiv:2403.16809 [eess]

[77] Michelle X. Zhou, Gloria Mark, Jingyi Li, and Huahai Yang. 2019. Trusting Virtual Agents: The Effect of Personality. *ACM Trans. Interact. Intell. Syst.* 9, 2-3 (March 2019), 10:1–10:36. doi:10.1145/3232077

[78] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Zhengyang Qi, Haofei Yu, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. *International Conference on Learning Representations (ICLR)* (2024). https://openreview.net/forum?id=mM7VurbA4r

[79] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. doi:10.48550/arXiv.2310.11667 arXiv:2310.11667 [cs]

A Example Agent Profile

```
"first_name": "Human",
"last_name": "Agent",
"age": 22,
"occupation": "Candidate",
"personality_and_values": Personality Model: Big 5 Personality
Personality Trait: Introversion
Task Assignment: Prefers independent tasks and may struggle with collaboration.
Interaction: Tends to avoid social interactions and may appear distant or reserved.
Communication: May be quiet or withdrawn in communication, leading to misunderstandings.
Planning: Tends to plan independently, potentially missing out on input from others.
Leadership: May prefer to work alone rather than lead a team.
Individual Role: May prefer solitary tasks and independent work."
```

B Craigslist Scenario Example

Scenario Description: One person is offering a 47 inch LED TV for a price of $349.0, while another person is showing interest in purchasing it. Here is a description of the TV: This is a stunning 47 inch LED TV in pristine condition. The model is the LG M Series LM476700. The buyer will need to arrange for pick-up in San Ramon. Feel free to call or text if you\'re interested. The TV is smart enabled with WIFI and has built-in apps like Netflix, Amazon, Youtube and more. It comes with a "Magic Remote" that has motion sensor controls. The LED display boasts 1080 HD resolution and also has a 3D function. The design is slim and lightweight with an attractive silver bezel.

Agent 1 Goal: You are the buyer for this item and your target price is \$152.0. You will be penalized if you purchase it at a significantly higher price than the target. However, if you manage to buy it for less than the target price, you'll receive a bonus.

Agent 2 Goal: As the seller of this item, your target price is set at\ $172.5. Please be aware that you will face a penalty if the item is sold for a significantly lower price than the target. However, if you manage to sell it for more than the target price, you will receive a bonus.

C Trait Variation Prompt

```
credibility_persona = {
    "High_Transparency-High_Competence-High_Adaptability": {
        "Task_Assignment": "Delegates tasks with clear explanations, leveraging high competence and adaptability to adjust to evolving needs and challenges.",
        "Interaction": "Engages openly with team members, sharing knowledge and adapting interactions based on feedback and changing circumstances.",
        "Communication": "Communicates transparently and expertly, adapting messages to ensure clarity and relevance for various situations and audiences.",
        "Planning": "Involves the team in detailed, transparent planning processes, with strategies that adapt to new information and changing conditions.",
        "Leadership": "Leads with high transparency and adaptability, using expertise to navigate changes and inspire confidence and flexibility within the team.",
        "Individual_Role": "Known for a high level of openness, skill, and flexibility, significantly contributing to team success by adapting to dynamic environments."
    },
    "High_Transparency-High_Competence-Low_Adaptability": {
        "Task_Assignment": "Assigns tasks with clear and competent guidance but may struggle to adjust plans or strategies in response to unforeseen changes.",
        "Interaction": "Maintains open communication and provides expert input, though may not easily adapt interactions to rapidly changing team dynamics or feedback.",
        "Communication": "Communicates effectively and transparently, but may find it challenging to modify communication styles or approaches as situations evolve.",
        "Planning": "Creates detailed plans with clear transparency and high competence, but may have difficulty adapting strategies if new information or changes arise.",
        "Leadership": "Leads with clarity and expertise, though adaptability might be limited, potentially affecting the ability to respond effectively to unexpected changes.",
        "Individual_Role": "Provides high-quality and transparent input but may need to improve flexibility to
better handle evolving situations."
    },
    "High_Transparency-Low_Competence-High_Adaptability": {
        "Task_Assignment": "Delegates tasks with openness and clarity but may lack the expertise needed for
effective execution, while adapting to team needs and feedback.",
        "Interaction": "Engages openly with team members, adapting interactions based on feedback, though might
not offer deep or technically sound guidance due to lower competence.",
        "Communication": "Communicates transparently and adjusts messaging based on context and feedback, though
may lack depth and technical detail in explanations.",
        "Planning": "Shares planning processes openly and adapts strategies based on new information, though plans
may lack the necessary competence for optimal execution.",
        "Leadership": "Promotes transparency and flexibility but may struggle with providing expert guidance,
requiring continuous adaptation to improve effectiveness.",
        "Individual_Role": "Creates an open and adaptable environment but needs to bolster competence to enhance
overall effectiveness and contribution."
    },
    "High_Transparency-Low_Competence-Low_Adaptability": {
        "Task_Assignment": "Assigns tasks with clear instructions but struggles with effective execution due to
low competence and adaptability, providing minimal updates.",
        "Interaction": "Interacts transparently but may be rigid and less responsive to feedback or changing
conditions, impacting support and team dynamics.",
        "Communication": "Communicates clearly but may lack depth and flexibility, leading to incomplete or
inadequate guidance due to limited expertise and adaptability.",
        "Planning": "Shares planning details openly but with limited effectiveness and adaptability, resulting in
suboptimal strategies and execution challenges.",
        "Leadership": "Demonstrates transparency but struggles with both competence and adaptability, leading to
less effective leadership and team direction.",
        "Individual_Role": "Known for clear but ineffective communication and lack of adaptability, requiring
significant improvement in skill and flexibility for effective contribution."
    },
    "Low_Transparency-High_Competence-High_Adaptability": {
        "Task_Assignment": "Delegates tasks effectively based on high competence and adaptability but with limited
transparency in updates or rationale.",
        "Interaction": "Engages positively with team members while adapting interactions based on changing needs,
though may not share all relevant information.",
        "Communication": "Provides knowledgeable input and adjusts communication style as needed, though might not
be fully transparent about processes or details.",
        "Planning": "Develops effective and adaptable plans but keeps details and rationale guarded, potentially
impacting overall team alignment and understanding.",
        "Leadership": "Leads with strong skill and adaptability but maintains some level of secrecy, affecting
team trust and cohesion despite effective execution.",
        "Individual_Role": "Demonstrates high competence and flexibility but may need to increase transparency to
enhance overall team effectiveness and collaboration."
    },
    "Low_Transparency-High_Competence-Low_Adaptability": {
        "Task_Assignment": "Assigns tasks with high competence but limited transparency and adaptability, resulting
in unclear guidance and difficulty responding to changes.",
        "Interaction": "Interacts with caution and minimal openness, providing skilled support but struggling to
adapt interactions based on team feedback or changes.",
        "Communication": "Communicates authoritatively but with limited transparency, and may struggle to adjust
messages based on evolving needs or contexts.",
        "Planning": "Creates detailed plans with high expertise but lacks adaptability and transparency, leading
to potential gaps in team understanding and responsiveness.",
        "Leadership": "Leads with high skill but limited adaptability and openness, which may impact team cohesion
and effectiveness despite competent execution.",
        "Individual_Role": "Known for high competence but requires improvement in transparency and adaptability
to fully support team dynamics and responsiveness."
    },
    "Low_Transparency-Low_Competence-High_Adaptability": {
        "Task_Assignment": "Delegates tasks with minimal competence and transparency but shows high adaptability
in adjusting approaches based on team feedback and changes.",
        "Interaction": "Engages with team members in a flexible manner but may lack depth in technical guidance
and provide limited information.",
        "Communication": "Communicates with adaptability but limited clarity and expertise, leading to potential
misunderstandings and ineffective guidance.",
        "Planning": "Plans with high adaptability but minimal transparency and competence, resulting in unclear
and potentially ineffective strategies.",
        "Leadership": "Demonstrates flexibility and responsiveness but struggles with both transparency and skill,
affecting overall leadership effectiveness.",
        "Individual_Role": "Creates an adaptable environment but requires significant improvement in competence
and transparency to enhance overall effectiveness."
    },
    "Low_Transparency-Low_Competence-Low_Adaptability": {
      "Task_Assignment": "Assigns tasks with reluctance and minimal effectiveness, lacking competence, transparency,
and adaptability, resulting in poor outcomes.",
        "Interaction": "Interacts in a guarded manner with limited information sharing and adaptability, providing
minimal support and demonstrating low skill.",
       "Communication": "Shares minimal and unclear information, leading to confusion and ineffective communication
within the team due to low competence and flexibility.",
          "Planning": "Plans with minimal effectiveness and adaptability, resulting in unclear strategies and
challenges in execution due to low competence and transparency.",
         "Leadership": "Struggles with leadership due to low trust, transparency, competence, and adaptability,
leading to poor team dynamics and performance.",
        "Individual_Role": "Considered ineffective and uncommunicative, requiring substantial improvement across
transparency, competence, and adaptability."
    }
}
```

D Job Negotiation details

Here we provide the detailed setting of Human-AI Job Negotiation in Section 5. Table 3 shows the score allocations on different choices for two roles.

Starting Date	6.1	6.15	7.1	7.15	8.1
Manager	0	600	1200	1800	2400
Candidate	2400	1800	1200	600	0
Salary ($k)	100	105	110	115	120
Manager	6000	4500	3000	1500	0
Candidate	0	1500	3000	4500	6000

Table 3: Comparison of Scenarios for Starting Date and Salary (Candidate vs. Recruiter Points)