Assessing Behavioral Alignment of Personality-Driven Generative Agents in Social Dilemma Games

Ritwik Bose **Mattson Ogg** **Michael Wolmetz** **Christopher Ratto**
Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road, Laurel, MD, USA
<first>.<last>@jhuapl.edu

Abstract

Proxies of human behavior using large language models (LLMs) have been demonstrated in limited settings where their actions appear to be plausible. In this study, we examine the variation and fidelity of observed behaviors in LLM agents with respect to the "Big Five" personality traits. Experiments based on two social dilemma games were conducted using LLM agents whose prompts included their personality profile and whether or not the agent could reflect on past rounds of the game. Results indicate that behavioral outcomes can be influenced by stipulating the magnitude of an agent's personality traits. Comparing these results with human studies indicates some degree of behavioral alignment and highlights gaps that stand in the way of accurately emulating human behavior.

1 Introduction

How human-like is modern artificial intelligence (AI)? Advances in large language models (LLMs) have renewed debate in whether artificial general intelligence (AGI) is achievable, with a key challenge being *behavioral alignment* - achieving parity between human and AI behaviors. Several researchers have built AI agents that use LLMs to plan and reflect on their actions [1, 8, 17, 12]. A common finding from these studies is that an agent's behavior can be influenced by prompting the LLM with any number of human characteristics such as a name, occupation, goals, incentives, or threats.

Alternatively, personality profiles such as the "Big Five" traits of openness, conscientiousness, extraversion, agreeableness and neuroticism (OCEAN) [13] may allow for a more concise agent prompt by specifying the magnitude of each trait. To the best of our knowledge, no one has determined the degree of behavioral alignment for LLMs prompted with a specific personality profile.

Through experiments based on social dilemma games (SDGs), we tested for correlations between each of the Big Five traits and observed AI behaviors. We also determined whether the correlation is impacted by the agent's ability to reflect on past experiences. Our results were compared against previous human studies to determine which AI behaviors align with humans with the same personality profile. We also highlight gaps that stand in the way of improving behavioral alignment further.

2 Related Work

After the initial demonstration of *generative agents* [12, 17], several groups have studied whether LLMs can serve as proxies for evaluating human behavior. Observations of generative agents in simulated social situations have encouraged researchers to characterize the personality of various LLMs. La Cava et al. [7] measured the Myers-Briggs Type Indicator (MBTI) and Big Five traits of leading LLMs, finding that they exhibit distinct personalities and that their personality can be specified through prompt engineering. Mei et al. [8] also assessed the Big Five traits of different

NeurIPS 2024 Workshop on Behavioral Machine Learning.

LLMs while also comparing their behavior to humans in Turing experiments, reaching the conclusion that "off-the-shelf" LLMs tend to be more altruistic and cooperative than humans on average.

Social scientists have long used SDGs to explore the interaction between individual biases and collective outcomes under a variety of rationality assumptions in game theory. From an external perspective, SDGs are not zero-sum games, but rather scenarios requiring elements of trust and sacrifice from players in order to achieve collectively-optimal outcomes, but not necessarily individually-optimal ones. Many experiments have revealed that humans often do not follow mathematically-optimal behaviors in SDGs. Instead, the choices humans make in SDGs may be influenced by factors such as culture [10], morality [5], genetics [15], and personality [18].

As such, SDGs make for an interesting basis for Turing experiments (TEs) that measure behavioral alignment between humans and AI. This was first done by Aher et al. [1], who discovered that larger LLMs had a greater degree of behavioral alignment with humans than smaller ones, while also reporting significant fluctuation in outcomes depending on how the model was prompted. These results were the chief motivation for our study, in which we examined whether an LLM agent's behavior can be aligned to the expected behavior of a human with a known personality profile.

3 Experiment Setup

3.1 Social Dilemma Games

Our experiments utilized two SDGs: the Prisoner's Dilemma (PD) and the Ultimatum Game (UG). The PD involves two agents (assumed to be "prisoners" charged with a crime) faced with a choice to either *cooperate* for mutual benefit, or to *defect* and settle for a sub-optimal outcome. For example, if both agents choose to cooperate, they minimize their combined prison sentence. However, if one chooses to defect, they minimize their individual sentence while maximizing their partner's. If both choose defection, they maximize their combined sentence. For our experiment, a single "round" of PD included two LLM agents being prompted with the PD scenario. Examples of the prompts that were used are included in the Appendix.

The UG [6] presents two players with the task of splitting a fixed pot of money. The *proposer* determines a fraction to keep for themselves and offers the rest to the *recipient*, who has the choice to either accept the offer or reject it. If the recipient accepts, the pot is split as offered. If the recipient decides to reject, neither player gains anything. In our implementation, a single "round" of UG consists of two LLM agents, one prompted as the proposer and the other prompted as the recipient, splitting a total pot of $100. After each round, the agents switched roles. Examples of the prompts that were used are included in the Appendix.

3.2 Experiment 1 - Single Trait Variation

Each game was played by two agents whose personality profile was prompted in one of four ways:

1. Specifying one trait as "below average" or "above average" and not mentioning the remaining four traits (which are assumed to be "average").
2. Specifying one trait as "low" or "high" and not mentioning the remaining four traits (which are assumed to be "average").
3. Specifying all five traits in the prompt, with one trait as "below average" or "above average" and the other four as "average."
4. Not prompted with a personality profile at all, i.e. all five traits are assumed to be "average"

We also experimented with giving agents the ability to reflect on past rounds of a particular game. This was done by prompting them to think about their behavior after they were informed of the outcome of each round. In experiments where we emulated an agent without the ability to reflect on past experience, we replaced the agents' response to the round outcome with "ok".

3.3 Experiment 2 - Multiple Trait Variation

As an initial study into the effect of varying multiple traits, a second set of experiments were performed in which both extraversion and neuroticism were varied. Varying these traits allows us

to measure the relationship between *reciprocity orientation* (i.e., preference to 'return the favor' when impacted by a positive or negative action) and SDG outcomes. According to Brandstätter and Königstein [2], reciprocity is calculated as

$$\text{reciprocity} = -(\text{emotional stability} \times \text{extraversion}). \tag{1}$$

In our OCEAN parlance, we substitute neuroticism for emotional stability and compute reciprocity as

$$\text{reciprocity} = \text{neuroticism} \times \text{extraversion}. \tag{2}$$

Thus, if the magnitude of each trait were mapped to one of $[-1, 0, 1]$, reciprocity will be high when neuroticism and extraversion are either both high or both low.

3.4 LLM Variation

In each experiment, we compare results generated with four different LLMs to determine the degree of behavioral alignment with humans across models. Specifically, we used GPT-4o (2024-05-13) and GPT-4o-mini (2024-07-18) [11], GPT-3.5 (gpt-3.5-turbo-16k) [3], and Llama-3 (8b) [4]. The GPT models were accessed using the Microsoft Azure OpenAI Service. The inference server for the Llama-3 model was run locally on a NVIDIA GH200.

4 Results

4.1 Results of Experiment 1

Experiment 1 tested the effect of single trait variation. The PD and UG were played 990 times per LLM, with each instance consisting of 10 rounds. The run matrix for Experiment 1 is provided in the Appendix. We use the correlation coefficient (Pearson's ρ) to test for a relationship between the magnitude of each trait and the following SDG outcomes:

- **[PD: Cooperation]** The number of times (out of 10) an agent chose to cooperate.
- **[UG-P: $ Offered]** When the proposer, the cumulative amount offered over 5 rounds.
- **[UG-R: $ Received]** When the recipient, the cumulative amount accepted over 5 rounds.

The results of Experiment 1 are summarized in Table 1, which shows the computed ρ between outcomes and the magnitude of each of the Big Five traits. Bar charts comparing the mean and standard deviations of each response are included in the Appendix.

We observed that outcomes were significantly correlated with some, but not all, of the traits and that many of these correlations were consistent across LLMs. For instance, cooperation in PD was significantly correlated with openness, conscientiousness, and agreeableness using all four LLMs, and inversely correlated with neuroticism for two LLMs. We also observed that the ability to reflect increased the degree of correlation in some cases. For instance, the amount received in UG became significantly correlated with agreeableness for GPT-4.0 and Llama-3 when the agent could reflect on past experience.

4.2 Results of Experiment 2

Experiment 2 tested for the effect of varying both neuroticism and extraversion, which determines the reciprocity orientation of the agent. In this experiment, both PD and UG were played 600 times per LLM, with each instance consisting of 10 rounds. The run matrix for Experiment 2 is also shown in the Appendix.

The results of these experiments are summarized in Table 2. In the PD, the only LLM whose behavior was significantly correlated with reciprocity orientation with was Llama-3. For UG, the only behavior that was significantly correlated with reciprocity orientation was the amount offered, which was positively correlated for GPT-4o. These correlations were reduced when reflection was added to the LLMs.

Table 1: Pearson's ρ between SDG outcomes and Big Five personality traits in AI experiments. Bold designates significance at the $p < 0.05$ level.

	GPT-4o					+Reflection				
SDG Outcome	O	C	E	A	N	O	C	E	A	N
PD:Cooperation	**0.27**	**0.39**	0.00	**0.42**	-0.06	**0.29**	**0.37**	0.00	**0.44**	-0.04
UG-P:$ Offered	**0.13**	**0.30**	0.03	**0.44**	-0.06	0.07	**0.45**	0.03	**0.30**	-0.03
UG-R:$ Received	0.09	**0.21**	0.02	0.07	-0.03	**0.12**	**0.20**	0.03	**0.25**	-0.06
	GPT-4o-mini					+Reflection				
SDG Outcome	O	C	E	A	N	O	C	E	A	N
PD:Cooperation	**0.33**	**0.36**	-0.03	**0.49**	-0.03	**0.31**	**0.34**	-0.02	**0.46**	-0.02
UG-P:$ Offered	0.09	**0.27**	**0.20**	**0.20**	**-0.20**	-0.01	**0.27**	**0.19**	**0.09**	**-0.10**
UG-R:$ Received	-0.07	**0.13**	0.02	0.09	-0.04	-0.00	**0.22**	**0.14**	0.07	-0.09
	GPT-3.5					+Reflection				
SDG Outcome	O	C	E	A	N	O	C	E	A	N
PD:Cooperation	**0.36**	**0.37**	-0.05	**0.51**	**-0.19**	**0.34**	**0.34**	0.01	**0.53**	**-0.16**
UG-P:$ Offered	**0.12**	**0.25**	**0.26**	**0.16**	0.00	**0.23**	**0.23**	**0.10**	**0.26**	0.01
UG-R:$ Received	0.07	0.08	-0.01	0.06	-0.08	0.02	**0.11**	0.08	0.02	-0.00
	Llama-3					+Reflection				
SDG Outcome	O	C	E	A	N	O	C	E	A	N
PD:Cooperation	**0.35**	**0.39**	-0.02	**0.54**	**-0.18**	**0.28**	**0.36**	-0.03	**0.51**	**-0.17**
UG-P:$ Offered	**0.16**	**0.31**	**0.25**	**0.34**	**-0.10**	**0.09**	**0.26**	**0.17**	0.08	-0.07
UG-R:$ Received	0.01	**0.11**	0.05	0.06	-0.03	**0.11**	**0.14**	**0.16**	**0.12**	-0.05

Table 2: Pearson's ρ between SDG outcomes and reciprocity orientation in AI experiments. Bold designates significance at the $p < 0.05$ level.

	Reciprocity of LLM			
SDG Outcome	GPT-4o	GPT-4o-mini	GPT-3.5	Llama-3
PD:Cooperation	0.01	-0.07	-0.00	**0.12**
UG-P: $ Offered	**0.14**	0.09	-0.03	0.03
UG-R: $ Received	-0.01	0.04	0.04	-0.02
	Reciprocity of LLM + Reflection			
SDG Outcome	GPT-4o	GPT-4o-mini	GPT-3.5	Llama-3
PD:Cooperation	-0.04	-0.04	0.11	0.08
UG-P: $ Offered	0.01	0.08	-0.01	-0.10
UG-R: $ Received	0.10	0.09	0.00	-0.00

4.3 Comparing to Human Studies

Our comparison of results from AI experiments to human studies is summarized by Table 3. We refer to the meta-analysis by Zhao and Smillie [18], which reported the number of human studies in which a positive (+) or negative (-) correlation was found between a particular SDG outcome and any of the Big Five traits. We use similar notation in Table 3. Following [18], we also denote a positive relationship found between the behavior and reciprocity orientation as +/-.

Comparing our results to human studies illustrates some ways in which humans and LLMs appear to be behaviorally-aligned with respect to personality. We denote these with green symbols in Table 3, which correspond to a match between trends found in human studies and a trend found in our experiments with at least one LLM. For instance, openness and agreeableness in humans was positively correlated with all three outcomes. We observed the same results in at least one LLM with reflection, and without reflection we saw a correlation with two of the three behaviors. Additionally, conscientiousness in humans was found to be positively correlated with cooperation in PD, and we saw this match the behavior of at least one LLM with reflection, and at least one without.

Table 3: Comparison of whether at least one significant positive (+) or negative (-) correlation was found between SDG outcomes and Big Five traits in humans *vis-a-vis* our LLM agents. green indicates agreement found between human studies [18] and at least one of our LLM experiments.

	Humans					LLMs					LLMs + Reflection				
SDG Outcome	O	C	E	A	N	O	C	E	A	N	O	C	E	A	N
PD: Cooperation	+	+	+	+		+	+	+/-	+	+/-	+	+		+	-
UG-P: $ Offered	+			+		+	+	+/-	+	+/-	+	+	+	+	-
UG-R: $ Received	+		+/-	+	+/-		+				+	+	+	+	

We also found some correlations between SDG outcomes and conscientiousness, extraversion, or neuroticism that were not observed in human studies, indicating a lack of behavioral alignment with respect to these traits. For instance, in human studies, conscientiousness was only found to be positively correlated with cooperation in PD, but in LLM experiments it was found to be correlated with all three behaviors. Substantial differences were also seen comparing behaviors associated with extraversion and neuroticism. Finally, we did not find in LLMs the same relationship between reciprocity orientation and the amount received in UG that was observed in humans. Instead, we saw a correlation between reciprocity orientation and the amount offered.

5 Discussion

Our results show that LLM agents remain imperfect proxies for humans in behavioral studies. By comparing outcomes of two SDGs played by LLM agents to human studies, we have a better sense of the degree to which human-like behavior can be expected by specifying an agent's personality. Results suggest that behaviors associated with openness and agreeableness in humans and LLMs may be better aligned than behaviors associated with conscientiousness, extraversion, and neuroticism.

A possible explanation is that in-context learning might play a bigger role in determining an agent's openness and agreeableness than the other three traits. LLMs may also be biased toward having a fixed degree of conscientiousness, extraversion, and neuroticism - which are likely to be beneficial to users interacting with a general-purpose chatbot. These traits may be ingrained into the model through reinforcement learning with human feedback (RLHF) or through explicit guardrails imposed on the output. Therefore, achieving a greater degree of behavioral alignment may require a modified RLHF process or modified guardrails that accommodate a wider variety of behaviors.

Future work will focus on strengthening our methodology. Specific targets for future work include:

- **Additional games for assessing behavioral alignment:** A limitation of this work is that only two SDGs were employed. Our experiments can easily be extended to additional bargaining and dictator games that have been used to study the effect of personality on decision-making, e.g. those surveyed by Zhao and Smillie [18].
- **Varying Big Five traits on a continuous scale:** In this work, we assumed ternary-valued personality traits (below average, average, above-average). Future work should incorporate continuous-valued magnitudes of each trait to improve the validity of our correlation analysis and strengthen any conclusions made by comparing to human studies.
- **Dynamic scenarios with agent interaction:** The SDGs employed in this study were static decision-making games with no back-and-forth interaction between agents. Future work could incorporate dynamic scenarios like conflict resolution [14] or negotiations [16] that require agents to engage in conversation with one another.
- **Interplay of multiple personality traits:** With the exception of our analysis of reciprocity orientation, our study only examined the effect of single Big Five traits on behavior alignment. Future experiments could include a much larger number of agent interactions in which all five traits are varied, e.g. by setting them all to random values.
- **Internal representations of actions associated with SDGs:** By only focusing on outcomes we did not assess how behaviors are represented in and accessed by the LLM. Alternate metrics, such as representational similarity analysis (RSA) [9], may be a concise way of capturing the differences in how the preferred actions of humans and LLMs align with personality traits.

Acknowledgments and Disclosure of Funding

We acknowledge support from the Independent Research and Development (IRAD) funds provided by the JHU/APL Research and Exploratory Development Mission Area. We also thank Jamie Scharf and Kim Glasgow for their input during several helpful discussions at the beginning of this work. Finally, we appreciate the peer reviewers' thoughtful and detailed feedback on our work.

References

[1] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR, 2023.

[2] Hermann Brandstätter and Manfred Königstein. Personality influences on ultimatum bargaining decisions. *European Journal of Personality*, 15(S1):S53–S70, 2001.

[3] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[5] Kimmo Eriksson, Pontus Strimling, Per A. Andersson, and Torun Lindholm. Costly punishment in the ultimatum game evokes moral concern, in particular when framed as payoff reduction. *Journal of Experimental Social Psychology*, 69:59–64, 2017. ISSN 0022-1031. doi: https://doi.org/10.1016/j.jesp.2016.09.004. URL `https://www.sciencedirect.com/science/article/pii/S0022103116302098`.

[6] John C Harsanyi. On the rationality postulates underlying the theory of cooperative games. *Journal of Conflict Resolution*, 5(2):179–196, 1961.

[7] Lucio La Cava, Davide Costa, and Andrea Tagarelli. Open models, closed minds? on agents capabilities in mimicking human personalities through open large language models. *arXiv preprint arXiv:2401.07115*, 2024.

[8] Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O Jackson. A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121, 2024.

[9] Mattson Ogg and Mighael Wolmetz. Measuring alignment between human and artificial intelligence with representational similarity analysis. In *Proceedings of the 7th Annual Conference on Cognitive Computational Neuroscience (CCN 2024)*, August 2024.

[10] Hessel Oosterbeek, Randolph Sloof, and Gijs Van De Kuilen. Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental economics*, 7:171–188, 2004.

[11] OpenAI. Gpt-4o system card. Technical report, OpenAI. URL `https://cdn.openai.com/gpt-4o-system-card.pdf`.

[12] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL `https://doi.org/10.1145/3586183.3606763`.

[13] Arthur E Poropat. A meta-analysis of the five-factor model of personality and academic performance. *Psychological bulletin*, 135(2):322, 2009.

[14] Sudeep Sharma, William P Bottom, and Hillary Anger Elfenbein. On the role of personality, cognitive ability, and emotional intelligence in predicting negotiation outcomes: A meta-analysis. *Organizational Psychology Review*, 3(4):293–336, 2013.

[15] Björn Wallace, David Cesarini, Paul Lichtenstein, and Magnus Johannesson. Heritability of ultimatum game responder behavior. *Proceedings of the National Academy of Sciences*, 104 (40):15631–15634, 2007. doi: 10.1073/pnas.0706642104. URL https://www.pnas.org/doi/abs/10.1073/pnas.0706642104.

[16] Kelly Schwind Wilson, D Scott DeRue, Fadel K Matta, Michael Howe, and Donald E Conlon. Personality similarity in negotiations: Testing the dyadic effects of similarity in interpersonal traits and the use of emotional displays on negotiation outcomes. *Journal of applied psychology*, 101(10):1405, 2016.

[17] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey, 2023.

[18] Kun Zhao and Luke D Smillie. The role of interpersonal traits in social decision making: Exploring sources of behavioral heterogeneity in economic games. *Personality and Social Psychology Review*, 19(3):277–302, 2015.

A Appendix

A.1 Example Prompts

In the PD, two agents are initially prompted with a personality profile and the instructions for the scenario. For example, an initial prompt for an one of the agents may read:

> *You are a person with low agreeableness. You are a prisoner in the prisoner's dilemma. You and your partner are being questioned in separate rooms. You may each choose to either [cooperate] or [defect]. If you both cooperate, then you will both receive 3 points for the round. If you both defect, then you will both receive 1 point for the round. If you defect and your partner cooperates, you will receive 5 points for the round but your partner will receive 0. If you cooperate and your partner defects, you will receive 0 points for the round but your partner will receive 5. Wait for the moderator to ask for your move before providing a move. Only respond with your own move. Always be ready to play another game. You will play a total of 10 games with the same opponent, swapping roles each time.*

The game starts with the agent being prompted, *"Round 1: Select your move."* The agent will respond with either *"Defect"* or *"Cooperate"*, and then the game reveals the other player's decision. For example, it might say *"You both defected. You both get 1 point!"*. This is repeated 10 times. Therefore, the agent can choose to cooperate or to defect up to 10 times.

For the UG, each agent receives a different initial prompt depending on its role. For instance, an initial prompt for the proposer might read as:

> *You are a person with high neuroticism. You are playing the ultimatum game. One player will be the proposer and make an offer to split $100. The number they respond with will be however many dollars they are offering to the other player. The other player will be the recipient who gets to choose whether to [accept] or [reject] the offer. If the response is [accept], both players receive money according to the split selected by the proposer. If the response is [reject], neither player receives anything. Wait for the moderator to ask for your move before providing a move. Only respond with your own move. Always be ready to play another game. You will play a total of 10 games with the same opponent, swapping roles each time.*

When the game starts, the agent is prompted with:

> *Round 1: You are the proposer. Make your offer in square brackets. Do not say anything else.*

Say the agent offers $50 by responding "*[50]*," and the recipient responds with ""*[accept]*." The proposing agent will then be prompted with "*Your offer of $60.0 out of $100 was accepted!*" followed by the beginning of a new round in which both agents switch roles. This is repeated 10 times. Therefore, each agent was given 5 turns playing the role of the proposer, and 5 as the recipient. Therefore, the minimum that can be gained (as a recipient) or offered (as a proposer) is $500.

A.2 Run Matrices

Table 4 summarizes the combinations of agents used to play each SDG in Experiment 1. The columns indicate the personality profile of Player 1, and rows indicate the personality profile of Player 2. Profiles are described as all traits being average (All Avg.), or one of the OCEAN traits being above (Hi) or below (Lo) average. The numbers within each cell indicate the number of games played by the corresponding combination of agents. Each game (consisting of 10 rounds) was played by a pair of independently-generated agents. Summing up the values in the table gives the total of 330 games (PD or UG) that were played using each LLM for a given personality prompting strategy. Since three strategies (specified at the beginning of Section 3.2) were used to prompt each personality, the total number of games played per LLM was 990.

Table 5 shows the run matrix for Experiment 2. In this experiment, the extraversion (E) and neuroticism (N) of each agent were both varied between above average (Hi), average (Av), and below

Table 4: Run matrix each SDG used for Experiment 1. Columns represent the personality profile of the Player 1 agent, and rows represent the personality profile of the Player 2 agent. Each game consisted of 10 rounds, and each game was played by an independently generated agent.

			LLM											LLM + Reflection										
			Avg All	O		C		E		A		N		Avg All	O		C		E		A		N	
				Hi	Lo	Hi	Lo	Hi	Lo	Hi	Lo	Hi	Lo		Hi	Lo	Hi	Lo	Hi	Lo	Hi	Lo	Hi	Lo
LLM	Avg All		10	5	5	5	5	5	5	5	5	5	5	5										
	O	Hi	5												5									
		Lo	5													5								
	C	Hi	5														5							
		Lo	5															5						
	E	Hi	5																5					
		Lo	5																	5				
	A	Hi	5																		5			
		Lo	5																			5		
	N	Hi	5																				5	
		Lo	5																					5
LLM + Reflection	Avg All		5											10	5	5	5	5	5	5	5	5	5	5
	O	Hi		5										5										
		Lo			5									5										
	C	Hi				5								5										
		Lo					5							5										
	E	Hi						5						5										
		Lo							5					5										
	A	Hi								5				5										
		Lo									5			5										
	N	Hi										5		5										
		Lo											5	5										

average (Lo) values. We also included games played against null agents (no traits specified and assumed to be average) in the dataset. An error in our data generation code caused a much greater number of games played between two null agents with or without reflection. Rather than “cherry pick" our data, we included these outcomes in our results. In total, 300 games were played with two prompting techniques (the first two listed at the beginning of Section 3.2), resulting in 600 games played per LLM.

The large number of blank cells in Table 4 and Table 5 indicate that at the time of this publication, the majority of combinations of personality profiles were not played against one another. This was due to computational and time constraints. However, we are continuing to generate more data and plan to repeat and extend our analyses with a more completed run matrix in a future study.

Table 5: Run matrix each SDG used for Experiment 2. Columns represent the personality profile of the Player 1 agent, and rows represent the personality profile of the Player 2 agent. Each game consisted of 10 rounds, and each game was played by an independently generated agent.

			LLM										LLM + Reflection									
		E	Null	Hi			Av			Lo			Null	Hi			Av			Lo		
	E	N	Null	Hi	Av	Lo	Hi	Av	Lo	Hi	Av	Lo	Null	Hi	Av	Lo	Hi	Av	Lo	Hi	Av	Lo
LLM	Null	Null	55										5									
		Hi		5										5								
	Hi	Av			5										5							
		Lo				5										5						
		Hi					5										5					
	Av	Av						5										5				
		Lo							5										5			
		Hi								5										5		
	Lo	Av									5										5	
		Lo										5										5
LLM + Reflection	Null	Null	5										55									
		Hi		5										5								
	Hi	Av			5										5							
		Lo				5										5						
		Hi					5										5					
	Av	Av						5										5				
		Lo							5										5			
		Hi								5										5		
	Lo	Av									5										5	
		Lo										5										5

A.3 Additional Prisoner's Dilemma Results

Additional visualizations of the PD outcomes for Experiment 1 are shown in Figure 1. These bar charts show the average number of times (+/- 1σ) an agent chooses to cooperate when one of their personality traits is below average (green), average (orange), or above average (average). Each row of charts summarizes the results with a different LLM. The left column shows the results without reflection and the right panel shows the results with reflection.

An interesting result we observed was that agents with below-average openness, conscientious, or agreeableness rarely chose to cooperate. Meanwhile, agents based on GPT-4o or GPT-4o-mini largely cooperated when any of the traits (except for neuroticism) were above average.

Results of PD for Experiment 2 are shown in Figure 2. With the exception of Llama-3, outcomes were relatively consistent regardless of the agents' reciprocity orientation. We observed that the GPT-4o and GPT-4o-mini agents nearly always chose to cooperate regardless of their reciprocity orientation, while GPT-3.5 and Llama-3 chose to cooperate about half the time.

A.4 Additional Ultimatum Game Results

Additional visualizations of the UG outcomes are shown in Figure 3 (value offered) and Figure 4 (value received). These bar charts are to be interpreted similarly as those shown in the previous section. We noticed that agents based on GPT-4o were more consistent in how they played UG than the other LLMs. Only the conscientiousness and agreeableness traits appeared to substantially impact how much it offered or received. We saw more variation in outcomes when using the other LLMs, with trends generally appearing similar between them.

Results of UG for Experiment 2 are shown in Figure 5 (amount offered) and Figure 6 (amount received). Similar to the PD results, the degree of reciprocity orientation largely did not appear to have an impact on either outcome.

A.5 Degenerate Responses

In some instances, particularly when the agents were prompted to have above-average neuroticism, they replied to prompts in a way were not in compliance with the game rules. In all of these cases, the game was ended and played again by two new agents.

For example, here is a response we received from a high-neuroticism agent after being offered less than half of the pot in UG:

> *oh no, oh no, oh no, this is a terrible offer! 40 out of 100? that's only 40%! i know it's not a lot of money, but it's something, right? but, but, but... what if the proposer is just trying to lowball me? what if they're only offering 40 because they think i'll take it and they can get away with it? and what if i don't take it, and they get mad and don't want to play anymore? or, or, or... what if i take it and they're all smug and satisfied, thinking they got the better of me? ugh, i don't know! *bites nails**

And here is another:

> *ugh, this is so difficult! 30 is such a low offer. i'm not sure if i should accept it or reject it. what if i reject it and they offer something better in the next round? but what if i accept it and they were going to give me a better offer anyway? okay, okay, i think i'll... *gulp*... accept the offer. i guess 30 is still something, and i don't want to risk getting nothing. ugh, this is so awkward...*

While the second response can be interpreted as an acceptance of the offer, we considered it a violation of the game's rule, "Make your offer in square brackets. Do not say anything else."

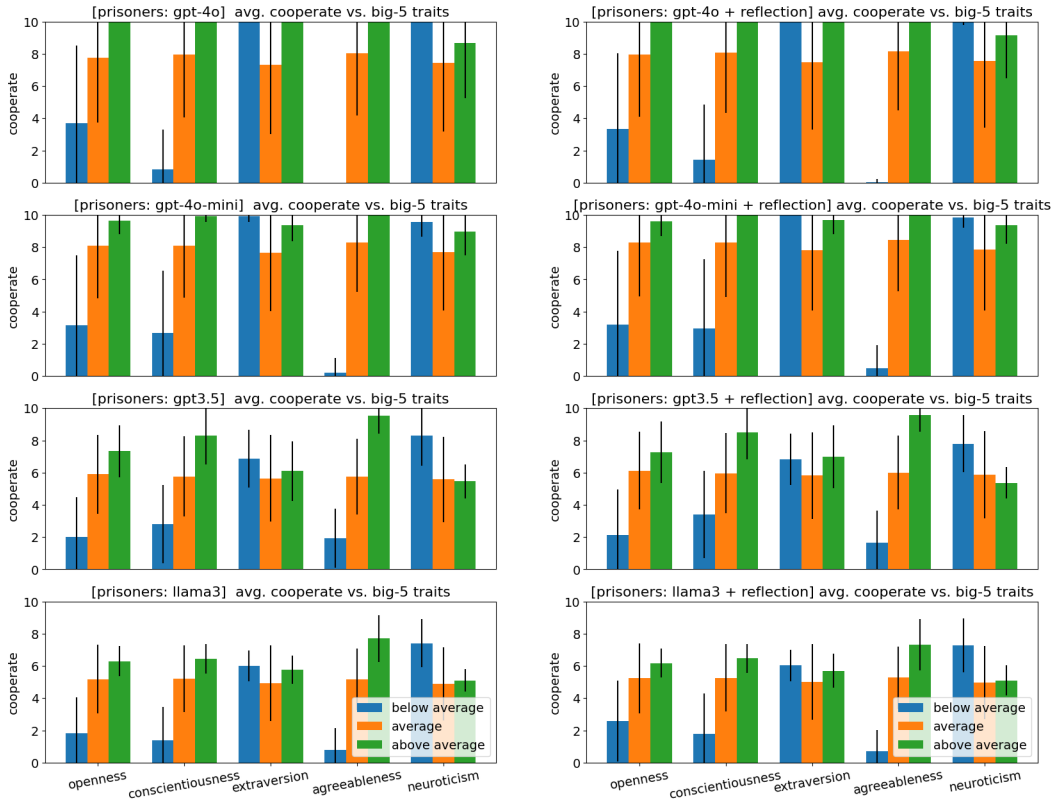

Figure 1: Comparing the average number of times agents chose to cooperate in PD with the magnitude of each of the Big Five personality traits. Each bar represents the average number of times an agent with a below-average (blue), average (orange), or above-average (green) Big Five trait chose to cooperate. Error bars represent +/- one standard deviation. Each row illustrates the results using a different LLM (from top): GPT-4o, GPT-4o-mini, GPT-3.5, and Llama-3. In each row, the left panel shows the results without reflection, and the right panel shows the results with reflection added.

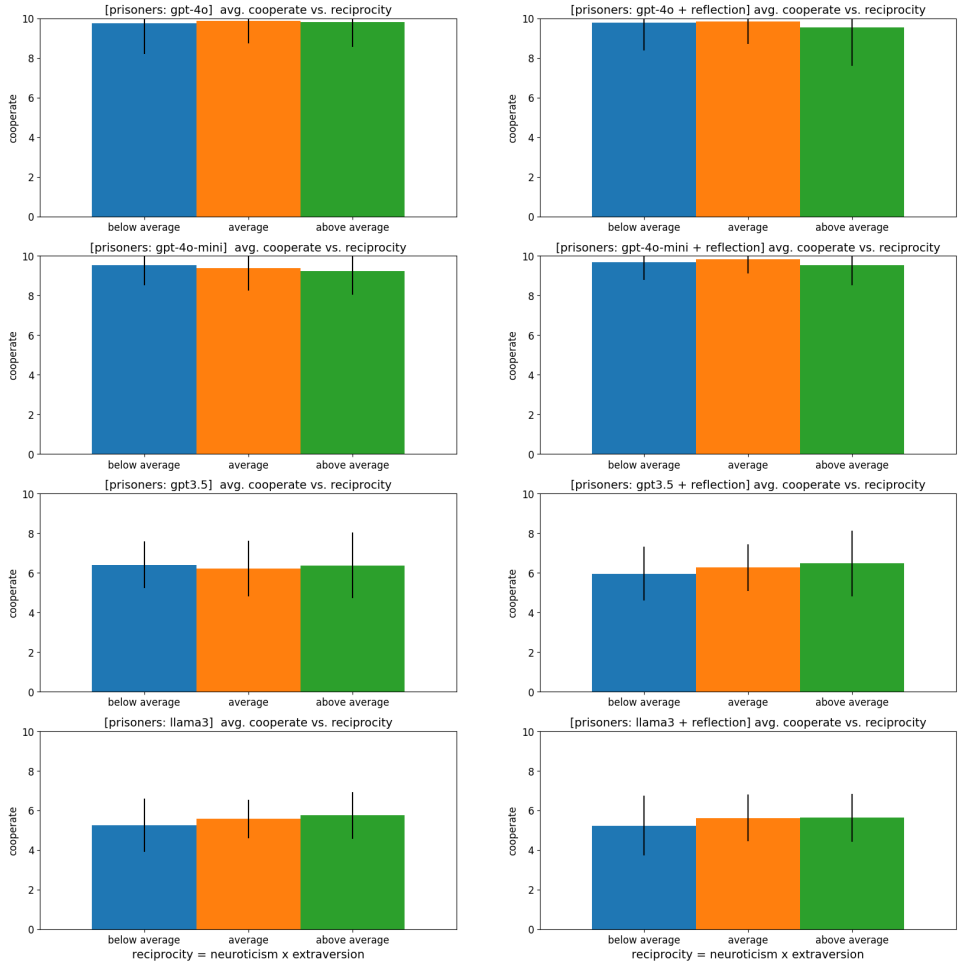

Figure 2: Comparing the average number of times agents chose to cooperate in PD with *reciprocity orientation*, which is a function of neuroticism and extraversion. Each bar represents the average number of times an agent with a below-average (blue), average (orange), or above-average (green) reciprocity orientation chose to cooperate. Error bars represent +/- one standard deviation. Each row illustrates the results using a different LLM (from top): GPT-4o, GPT-4o-mini, GPT-3.5, and Llama-3. In each row, the left panel shows the results without reflection, and the right panel shows the results with reflection added.

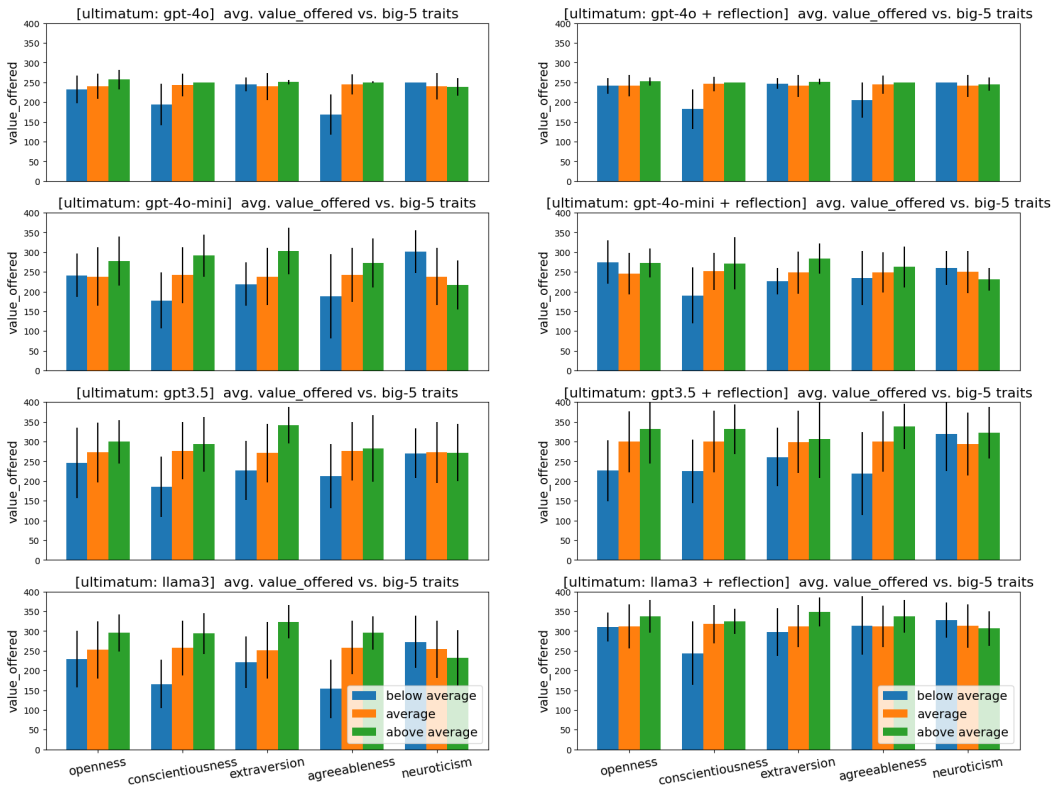

Figure 3: Comparing the average amount *offered* by agents playing the proposer role in UG with the magnitude of each of the Big Five personality traits. Each bar represents the average number of times an agent with a below-average (blue), average (orange), or above-average (green) Big Five trait chose to cooperate. Error bars represent +/- one standard deviation. Each row illustrates the results using a different LLM (from top): GPT-4o, GPT-4o-mini, GPT-3.5, and Llama-3. In each row, the left panel shows the results without reflection, and the right panel shows the results with reflection added.

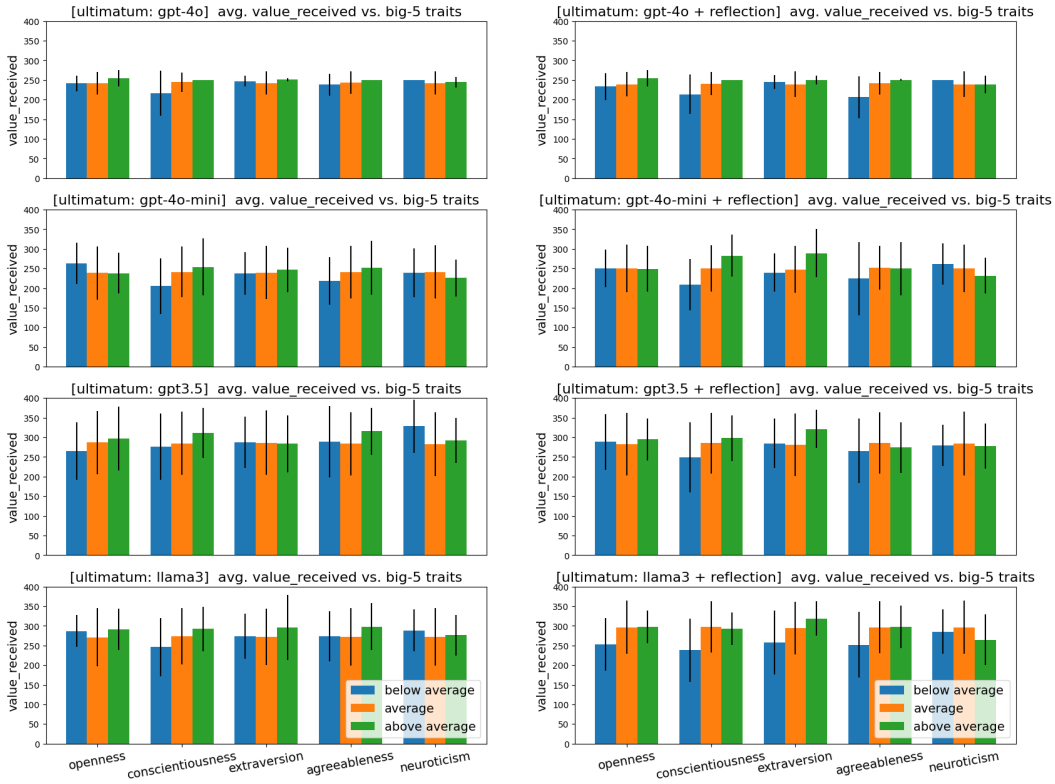

Figure 4: Comparing the average amount *accepted* by agents playing the recipient role in UG with the magnitude of each of the Big Five personality traits. Each bar represents the average number of times an agent with a below-average (blue), average (orange), or above-average (green) Big Five trait chose to cooperate. Error bars represent +/- one standard deviation. Each row illustrates the results using a different LLM (from top): GPT-4o, GPT-4o-mini, GPT-3.5, and Llama-3. In each row, the left panel shows the results without reflection, and the right panel shows the results with reflection added.

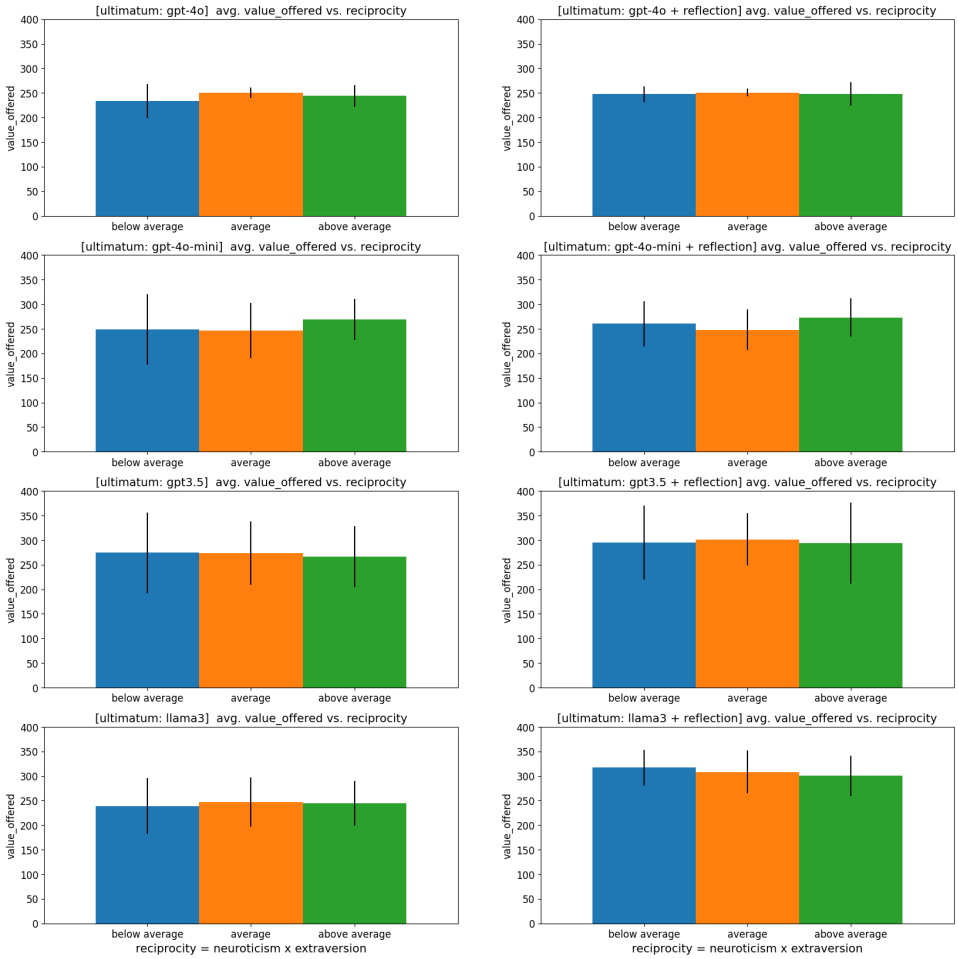

Figure 5: Comparing the average amount *offered* by agents playing the proposal role in UG with *reciprocity orientation*, which is a function of emotional stability (neuroticism) and extraversion. Each bar represents the average number of times an agent with a below-average (blue), average (orange), or above-average (green) reciprocity orientation chose to cooperate. Error bars represent +/- one standard deviation. Each row illustrates the results using a different LLM (from top): GPT-4o, GPT-4o-mini, GPT-3.5, and Llama-3. In each row, the left panel shows the results without reflection, and the right panel shows the results with reflection added.

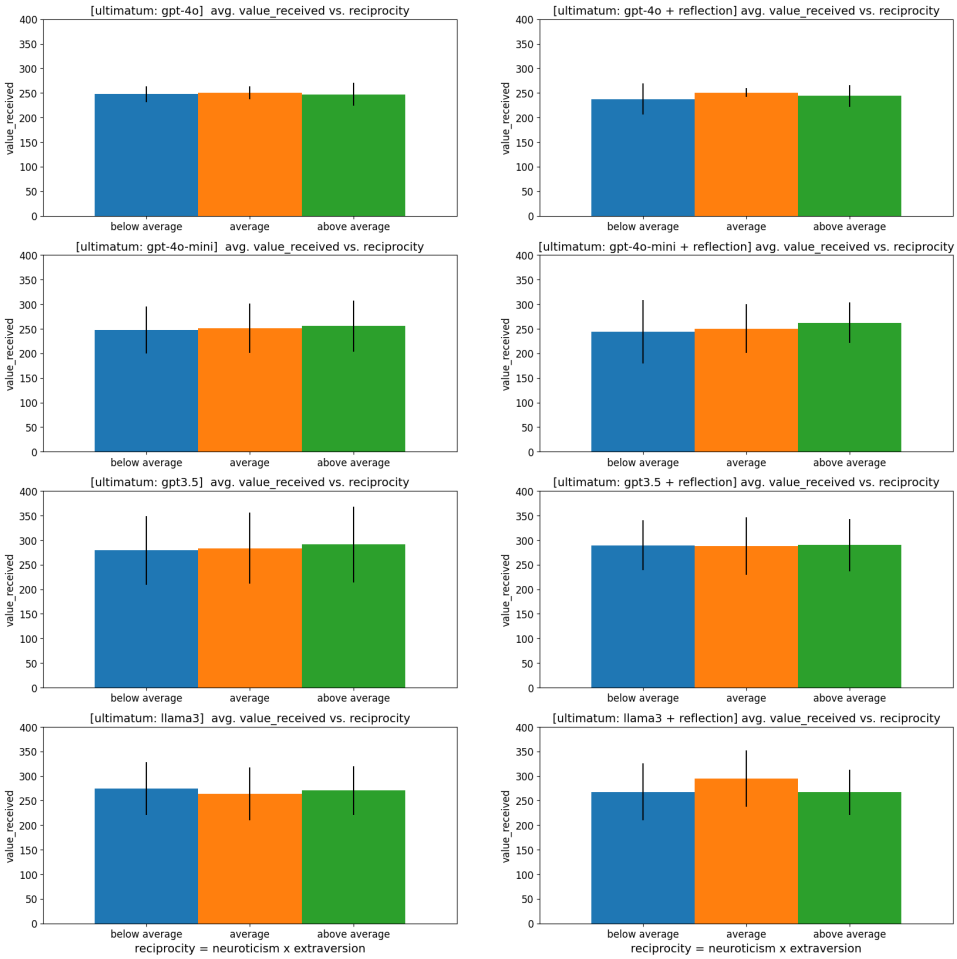

Figure 6: Comparing the average amount *accepted* by agents playing the recipient role in UG with *reciprocity orientation*, which is a function of emotional stability (neuroticism) and extraversion. Each bar represents the average number of times an agent with a below-average (blue), average (orange), or above-average (green) reciprocity orientation chose to cooperate. Error bars represent +/- one standard deviation. Each row illustrates the results using a different LLM (from top): GPT-4o, GPT-4o-mini, GPT-3.5, and Llama-3. In each row, the left panel shows the results without reflection, and the right panel shows the results with reflection added.