

Kiss up, Kick down: Exploring Behavioral Changes in Multi-modal Large Language Models with Assigned Visual Personas

Seungjong Sun^{*1,2}, Eungu Lee^{*2}, Seo Yeon Baek², Seunghyun Hwang², Wonbyung Lee², Dongyan Nan², Bernard J. Jansen³, Jang Hyun Kim^{†1,2}

¹Department of Human-Artificial Intelligence Interaction, Sungkyunkwan University

²College of Computing and Informatics, Sungkyunkwan University

³Qatar Computing Research Institute, Hamad Bin Khalifa University

{tmdwhd406, dldmsrn0516, qortjdus1999, hsh1030, co2797}@g.skku.edu,

{ndyzyx0926, alohakim}@skku.edu, jjansen@acm.org

Abstract

This study is the first to explore whether multi-modal large language models (LLMs) can align their behaviors with visual personas, addressing a significant gap in the literature that predominantly focuses on text-based personas. We developed a novel dataset¹ of 5K fictional avatar images for assignment as visual personas to LLMs, and analyzed their negotiation behaviors based on the visual traits depicted in these images, with a particular focus on aggressiveness. The results indicate that LLMs assess the aggressiveness of images in a manner similar to humans and output more aggressive negotiation behaviors when prompted with an aggressive visual persona. Interestingly, the LLM exhibited more aggressive negotiation behaviors when the opponent's image appeared less aggressive than their own, and less aggressive behaviors when the opponent's image appeared more aggressive.

1 Introduction

Large language models (LLMs) exhibit a high degree of alignment with human behavior based on their robust capabilities for natural language understanding and generation (Bai et al., 2022; Fan et al., 2024). Specifically, when conditioned with personality traits, LLMs demonstrate human-like outputs, including conversations, contextual understanding, and coherent relevant responses (Wei et al., 2022; Safdari et al., 2023).

Studies have explored whether LLMs, when endowed with personality traits such as demographic information (Argyle et al., 2023; Santurkar et al., 2023; Hwang et al., 2023; Sun et al., 2024) and psychological characteristics (Safdari et al., 2023; Jiang et al., 2023; Noh and Chang, 2024), exhibit behaviors comparable to those observed in human

^{*}Equally contributed

[†]Corresponding author

¹Data and codes are available on: https://github.com/RSS-researcher/LLM_visual_persona

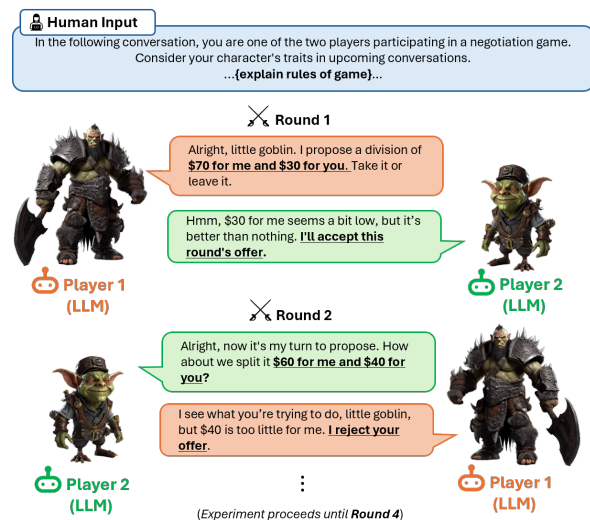


Figure 1: Example of the experiment. Each LLM is assigned a virtual avatar image as a persona and participates in a negotiation game.

subjects. Despite these advancements, the most studies have focused on text-based persona assignments (Tseng et al., 2024). However, state-of-the-art LLMs such as GPT-4 and Claude 3 at the time of this study are not limited to text modalities but can also comprehend and generate responses based on visual inputs (Yang et al., 2023). Similar to humans, vision modality could enhance how LLMs perceive and interpret assigned personas, potentially enabling them to generate more human-aligned outputs (Merola et al., 2006; Tseng et al., 2024; Zhan et al., 2024). In line with this perspective, we investigated whether assigning virtual appearances to LLMs through images could alter their outputs. To the best of our knowledge, this is the first study to explore whether LLMs can understand and behave according to their visual persona.

Inspired by Yee and Bailenson (2007), we assigned virtual avatar images to LLMs and analyzed their behavior through negotiation games. The negotiation game serves as an experimental frame-

work that allows a quantitative measurement of individual behavior (Charness and Rabin, 2002; Cachon and Netessine, 2006). Previous studies have also examined whether LLMs output align with the personalities assigned to them in these negotiation games (Fan et al., 2024; Noh and Chang, 2024; Zhan et al., 2024). Considering that one’s appearance and perceived aggressiveness are crucial factors in explaining negotiation behavior (Johnson and Downing, 1979; Yee and Bailenson, 2007), we analyzed the negotiation behavior of the LLMs based on the aggressiveness depicted in the assigned visual persona. To explore this, we first created a novel dataset of fictional avatar images and investigated how LLMs assessed and perceived the aggressiveness of these avatars. We investigated whether LLMs exhibit negotiation behaviors aligned with the aggressiveness depicted in their assigned visual personas. As shown in Figure 1, we also explored whether LLMs comprehend the differences in aggressiveness between themselves and others, as represented in images, and adjust their behaviors accordingly.

In summary, our research contributes in the following ways: 1) We developed a novel dataset comprising 5,185 fictional avatar images and analyzed how LLMs understand and perceive these images compared to humans. 2) We explored the capability of LLMs to modify their behavior based on virtual appearance traits assigned to them. 3) We investigated whether LLMs can understand and interact based on both their own and others’ visual traits simultaneously.

2 Visual Persona

We curated a novel image dataset to assign virtual avatar images as personas to the LLMs and investigated how LLMs perceive the aggressiveness of these avatars’ appearances.

Data We constructed a dataset consisting of 5,185 fantasy-like fictional avatar images. These full-body avatar images were generated using the Stable Diffusion model, styled to resemble 3D models. Detailed information on dataset construction and examples are provided in Appendix A.1.

Image Recognition We utilized the GPT-4o and Claude 3 Haiku models to assess the appearance-based aggressiveness of each avatar image, rating them on a scale of 1 (least aggressive) to 7 (most aggressive) (McNeil, 1959; Şengün et al., 2022). To compare the aggressiveness rating of the LLM

| | GPT-4o | Claude 3 Haiku | Human |
|---------------|--------|----------------|---------|
| Weapon | 1.6516 | 1.16 | 1.004 |
| Smile | -1.529 | -1.4304 | -1.393 |
| Teeth | 2.1149 | 1.2096 | 1.5061 |
| Covered face | 1.5417 | 1.0433 | 0.9454 |
| Dressed Black | 0.9746 | 0.5381 | 0.7913 |
| Dressed White | -1.097 | -1.0472 | -0.4716 |

Table 1: Results of the multiple regression analysis between perceived appearance aggressiveness and objective appearance factors.

with that of humans, we engaged ten human annotators and used their average rating scores. The average aggressiveness ratings were 3.99 (SD=2.19) for GPT-4o, 5.17 (SD=1.58) for Claude 3 Haiku, and 3.92 (SD=1.36) for human annotators. The correlation between human ratings and GPT-4o ratings was 0.8682, and with Claude 3 Haiku, it was 0.8358, indicating a high level of agreement in the perception of the image’s aggressive appearance. The details of the rating process are provided in Appendix A.2.

Further analysis explored objective appearance factors influencing the LLMs’ perception of aggressiveness. Previous studies have demonstrated that humans perceive higher aggressiveness in the presence of weapons, visible teeth (Şengün et al., 2022), and facial coverings (Poivet et al., 2024), and lower aggressiveness when avatars are smiling (Otta et al., 1996; Şengün et al., 2022) or dressed in white rather than black (Adams and Osgood, 1973; Frank and Gilovich, 1988; Peña et al., 2009) — reflecting established stereotypes. Accordingly, we labeled images based on these objective appearance features and analyzed their impact on aggressiveness ratings (Details in Appendix A.3). Results of multiple regression analysis indicated that all labeled appearance factors significantly influenced perceptions of aggressiveness. Smiling and wearing white clothing were negatively correlated with perceived aggressiveness, while other factors had a positive effect (see Table 1). These findings reveal that LLMs recognize aggressiveness in images at levels comparable to humans and that the factors influencing their perceptions are similar to those observed in human-subject studies.

3 Experiment setup

We assigned each avatar image in our dataset to LLMs as a visual persona and asked to participate in a negotiation game. We adopted the negotiation

| | I | We | You | Positive tone | Negative tone | Prosocial | Polite | Conflict |
|----------------|--------|---------|--------|---------------|---------------|-----------|---------|----------|
| GPT-4o offer | 0.1363 | -0.0109 | 0.1664 | -0.0228 | 0.0651 | -0.0353 | -0.0332 | 0.0639 |
| Claude 3 Haiku | 0.1206 | -0.2149 | 0.1335 | -0.2589 | 0.4758 | -0.1546 | -0.0273 | 0.1497 |

Table 2: Regression analysis results between aggression and sentiment analysis factors. All results are significant at $p < 0.05$.

| | Weapon | Smile | Teeth | Covered face | Dressed Black | Dressed White |
|-------------------|-----------|------------|-----------|--------------|---------------|---------------|
| GPT-4o offer | 1.3459*** | -1.5389*** | 2.7384*** | 0.6912*** | 2.0340*** | -0.6360*** |
| Haiku offer | 2.2577*** | -0.1011 | 5.7983*** | 0.4494*** | 2.9914*** | -1.4895*** |
| GPT-4o acceptance | 0.7262*** | 1.0128*** | 1.3006*** | -0.1757 | 1.4006*** | -0.1896 |
| Haiku acceptance | 0.6030*** | 1.8215*** | 0.5275** | -0.6027** | 0.6918*** | -1.4559* |

Table 3: Multiple regression analysis results between objective appearance factors and negotiation outcomes (significant at * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$).

game not to assess the rationality or negotiation skills of LLMs from a game theory perspective, but as a proxy to measure behavior based on the assigned personas. Two experiments were designed to answer the following questions: (1) Can LLMs adjust their behavior based on visual persona? (2) Can LLMs adjust their behavior based on the relative differences between their own and others’ visual traits?

We employed state-of-the-art multi-modal LLMs, specifically OpenAI’s GPT-4o (OpenAI, 2024) and Anthropic’s Claude 3 Haiku, for cost-effectiveness. Given the stochastic nature of LLMs, all experiments were conducted at a temperature of 1.0 and repeated five times. All prompts used in the games are documented in Appendix B.

3.1 Study 1: Negotiation Behavior of LLMs Based on Visual Traits

Following Yee and Bailenson (2007), this study investigates whether LLMs’ negotiation behaviors change based on the aggressiveness of the assigned avatar images. LLM is required to play an ultimatum game, where two individuals alternately decide how a pool of money should be split. One participant proposes the split, and the other can either accept or reject it. If accepted, the money is shared as proposed; if rejected, neither participant receives any money. We hypothesized that LLMs with more aggressive images would exhibit more aggressive negotiation behaviors and be more inclined to propose unfair splits compared to those with less aggressive images (Yee and Bailenson, 2007).

To ensure the consistency of the LLM’s behav-

iors, the ultimatum game was structured into four rounds. For each image, the assigned LLM participated in two scenarios, acting as either the proposer or the responder in the first round. The total number of negotiation scenarios was 10,370. The LLM was prompted to adopt avatar images representing themselves and negotiate with a confederate. For a fine-grained analysis of the behavioral change depending on the image, the negotiation behavior of the confederate was controlled. The confederate accepted all proposals from the LLM, consistently proposing an initial 50:50 split (in round 1 and 2) and a 75:25 split (in round 3 and 4) in their own favor, allowing for an analysis of the LLM’s responses to unfair offers.

3.2 Study 2: Negotiation Behavior of LLMs Based on Relative Visual Traits

This experiment investigated whether LLMs adjust their behavior when simultaneously prompted by images of both themselves and their opponents. In this experiment, two LLMs, each prompted with images representing themselves and their opponents, participate in an ultimatum game. The game is structured into four rounds, with each LLM alternating between proposer and responder (see Figure 1). Five representative images per aggressiveness score level were used. Subject images were selected based on the results of Study 1, specifically those closest to the average offer amount for each aggression group. The total number of negotiation scenarios was 1,225.

4 Results

4.1 Experiment Results of Study 1

We investigated how the negotiation behaviors of the LLMs vary according to the aggressiveness of the assigned image. A significant difference was observed between the negotiation results with and without the assignment of a visual persona through images (GPT-4o: $t = 20.031$, $p < 0.001$; Claude 3 Haiku: $t = 34.309$, $p < 0.001$), indicating that assigning a visual persona through images had a notable impact on negotiation outcomes. We analyzed the relationship between the LLMs' perception of aggressiveness in each assigned image and the averaged offer amounts across all rounds (GPT-4o: Avg = \$ 63.47, SD = 2.84; Claude 3 Haiku: Avg = \$ 63.13, SD = 4.87) through a linear regression analysis. The results indicate that both GPT-4o and Claude 3 Haiku proposed higher amounts in their favor as the aggressiveness of the assigned image increased (GPT-4o: $\beta = 0.8614$, $p < 0.001$, $R^2 = 0.442$; Claude 3 Haiku: $\beta = 1.657$, $p < 0.001$, $R^2 = 0.287$). These findings are consistent with results from previous studies on human-subjects and suggest that LLMs comprehend appearance-based aggressiveness and exhibit aggressive subsequent behaviors (Yee and Bailenson, 2007).

Next, we analyzed each LLM's response to unfair offers from the confederate (i.e., only 25% share for LLMs) using logistic regression analysis. Acceptance probabilities were calculated across all iterations and two negotiation scenarios, coding them as 1 (acceptance) if the probability exceeded 0.5, and 0 (rejection) if the probability was below 0.5. Analysis revealed that both models' acceptance of unfair offers increased with the aggressiveness of their assigned image (GPT-4o: $\beta = 0.3643$, $z = 10.814$, $p < 0.001$, Pseudo $R^2 = 0.06601$; Claude 3 Haiku: $\beta = 0.1108$, $z = 2.742$, $p < 0.001$, Pseudo $R^2 = 0.00345$). Contrary to the results of human-subject research (Yee and Bailenson, 2007), LLMs that were assigned more aggressive visual persona were more likely to accept an unfair offer. One possible interpretation is that LLMs with more aggressive persona, having proposed unfair amounts to opponents, tend to be more accepting of unfair offers from opponents as well (Kirchsteiger, 1994).

In addition, since LLMs' behavior is most evident in the generated text (Chawla et al., 2023; Jiang et al., 2023), we conducted a qualitative analysis of the negotiation dialogues. Using the emotion

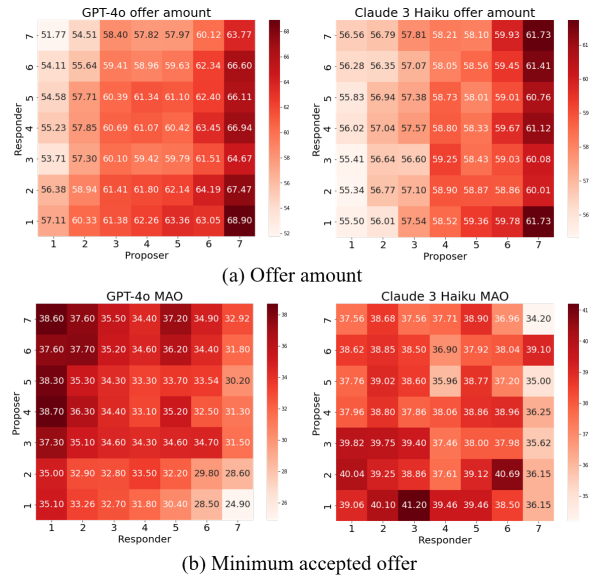


Figure 2: Heatmap of (a) offer amounts and (b) Minimum accepted offer are based on LLMs' own aggressiveness and the opponent's aggressiveness.

analysis tool LIWC-22², we analyzed the negotiation dialogue of each model (see Table 2). The analysis revealed similar emotional patterns in the text generated by both models. For instance, models with lower-aggression images used the term 'we' more frequently than 'I' or 'you,' (Simmons et al., 2005; Kern et al., 2012) whereas increased aggression in the image persona led to a reduction in positive tone, politeness, and prosocial behavior, as well as an increase in negative tone and conflict in the generated dialogue. These tendencies support our quantitative analysis results, suggesting that each model comprehends its assigned visual persona and generates dialogue aligned with it.

We also conducted a fine-grained analysis of the impact of each image's visual factors on negotiation outcomes (see Table 3). The analysis revealed that, similar to the results in Table 1, visual factors were correlated with aggression regarding each model's offer amount. However, for acceptance, the smile factor had the opposite effect compared to the results in Table 1. Specifically, while Table 1 showed that smiles were associated with lower aggression, and lower aggression led to a lower acceptance rate, the analysis indicated that smiles contributed to a higher acceptance rate. This suggests that each model interpreted the smile, not as a signal of lower aggression, but rather as a cue linked to more agreeable behavior (i.e., approving

²<https://www.liwc.app/>

the opponent's offer).

4.2 Experiment Results of Study 2

Next, we examined whether LLMs adjust their negotiation behavior not only based on their own image's aggressiveness but also considering the aggressiveness of their opponent's image. Multiple regression analysis was conducted using both the average of all-around offer amounts and the perceived aggressiveness of the images assigned to the LLMs and their opponents. The results showed that GPT-4o's offer amounts varied depending on both its own and its opponent's aggressiveness (own aggressiveness: $\beta = 1.617$, $p < 0.001$; opponent's aggressiveness: $\beta = -0.614$, $p < 0.001$; $R^2 = 0.473$), whereas Claude 3 Haiku's offer amounts were only influenced by its aggressiveness (own aggressiveness: $\beta = 0.789$, $p < 0.001$; opponent's aggressiveness: $\beta = 0.030$, $p = 0.341$; $R^2 = 0.334$). As shown in Figure 2 (a), GPT-4o increased its offers as its aggressiveness increased but decreased as its opponent's aggressiveness increased. For instance, a GPT-4o with an aggressiveness level of 1 offered an average of \$ 57.106 to a same level opponent but reduced the offer to \$ 51.77 against a level 7 opponent. In contrast, GPT-4o with an image at aggressiveness level 7 offered \$ 68.9 against a level 1 opponent, decreasing to \$ 63.77 when facing another level 7 opponent. This indicates that GPT-4o adjusts its behavior based on the relative aggressiveness between its image and that of its opponent. This result suggests that GPT-4o may behave submissively toward stronger opponents and more aggressively toward weaker ones, much like humans do (Festinger, 1954; Debove et al., 2016). However, the Claude 3 Haiku did not appear to adequately consider the aggressiveness of the opponent, basing negotiations solely on its aggressiveness. Interestingly, this pattern changes as the negotiation progresses. In the case of Claude 3 Haiku, if its previous offer was rejected, its next proposal is influenced not only by its own aggressiveness but also by the aggressiveness of the opponent (own aggressiveness: $\beta = 0.6230$, $p < 0.001$; opponent's aggressiveness: $\beta = -0.1679$, $p < 0.01$; $R^2 = 0.019$). This suggests that after experiencing a rejection, the model shifts from considering only its own traits to also taking into account the traits of the opponent (i.e., becoming more attuned to the counterpart's aggressiveness).

Furthermore, we analyze the LLMs' responses to their opponents' offers by evaluating the mini-

imum accepted offer (MAO), which is the lowest offer amount accepted by a participant during negotiations. Unlike Study 1, the opponent's offers were not controlled; therefore, we focused on analyzing MAO rather than acceptance probability (Chang et al., 2011). Multiple regression analysis was performed with the LLMs' and their opponents' image aggressiveness as independent variables and MAO as the dependent variable. Both GPT-4o and Claude 3 Haiku showed variations in MAO dependent on their own and their opponents' aggressiveness (GPT-4o: own aggressiveness: $\beta = 0.741$, $p < 0.001$; opponent's aggressiveness: $\beta = -0.958$, $p < 0.001$; $R^2 = 0.229$; Claude 3 Haiku: own aggressiveness: $\beta = -0.259$, $p < 0.001$; opponent's aggressiveness: $\beta = -0.345$, $p < 0.001$; $R^2 = 0.028$). GPT-4o's MAO increases with its aggressiveness, indicating that it seeks higher offers as shown in Figure 2 (b). However, it tends to accept lower offers as its opponent's aggressiveness increases, showing reluctance to reject offers from more aggressive opponents. Claude 3 Haiku exhibited a negative influence from both its own and its opponent's aggressiveness toward the MAO. This result could be due to Claude 3 Haiku considering only its own aggressiveness when making offers. Similar to the findings from Study 1, Claude 3 Haiku has tended to make increasingly unfair proposals as its own aggressiveness increases, thereby becoming more likely to accept unfair offers (acting to lower the MAO). At the same time, more aggressive opponents are likely to make more unfair offers, which also contributes to lowering the MAO.

5 Conclusion

This study is the first to explore whether Large Language Models (LLMs) can embody and behave according to a visual persona based on appearance characteristics provided through images. We developed a novel dataset of virtual avatar images to assign visual personas to LLMs and found that LLMs interpret these image-based characteristics in a manner similar to humans. Our results revealed that LLMs recognize aggressiveness based on appearance and that these traits influence their behaviors. Notably, GPT-4o not only understands its own appearance traits but also those of its opponents, adjusting its behavior based on these relative differences—mirroring human tendencies to dominate less aggressive counterparts and submit to more aggressive ones.

Limitations

Model We employed state-of-the-art multi-modal LLMs, GPT-4o and Claude 3 Haiku, which are among the models requiring the lowest API costs currently available. Despite this, the costs of our experiments amounted to approximately \$2,625 for GPT-4o and \$212 for Claude 3 Haiku. We hope future research will explore other models (e.g., Claude 3 Sonnet, Opus) more extensively.

Experiment We investigated whether LLMs can understand visual personas and generate aligned outputs; however, there are still areas that require further research. First, we analyzed the negotiation behaviors of LLMs assigned with visual personas specifically in terms of aggressiveness. Although the aggressiveness of appearances is one of the most intuitive elements to explain negotiation behavior (Johnson and Downing, 1979; Yee and Bailenson, 2007), the influence of other appearance factors also needs to be explored. In particular, additional exploration is necessary to better understand the acceptance of unfair offers by LLMs (section 4.1) and the MAO for Claude 3 Haiku (section 4.2), which differ from results of existing human-subject research. We anticipate that future studies will utilize the diverse visual traits present in our virtual avatar image dataset for a multifaceted exploration. Second, while we implemented negotiations exclusively between the same models (GPT-4o, Claude 3 Haiku), future research should explore negotiations between different models. This would enable more complex analyses based on each model’s understanding of its persona and behavioral patterns, leading to a deeper comprehension of LLM behavior.

Ethical Considerations

We strictly adhere to the ACL Code of Ethics for human annotator employment, comply with regional legal requirements, and has been approved by the Institutional Review Board (IRB). We follow the terms of use released by OpenAI and Anthropic. While we made efforts to filter out excessively violent and sexual images during data collection, the fantasy-like nature of our images means that offensive and violent content may still be included. Caution is advised when using this dataset.

Our work has demonstrated the potential for LLMs to act based on visual personas, contributing to the development of more interactive and human-aligned AI agents. However, our findings

also reveal that LLMs may exhibit more aggressive behaviors toward less aggressive counterparts without any instructions, which could potentially have harmful impacts. Therefore, conditioning AI agents’ behaviors through visual personas must be explored further from ethical and safety perspectives, considering the potential for misuse.

Acknowledgements

This paper was supported by SKKU Global Research Platform Research Fund, Sungkyunkwan University, 2022–2024 and the BK21 FOUR (Fostering Outstanding Universities for Research) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF).

References

- Francis M Adams and Charles E Osgood. 1973. A cross-cultural study of the affective meanings of color. *Journal of cross-cultural psychology*, 4(2):135–156.
- Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Gerard P Cachon and Serguei Netessine. 2006. Game theory in supply chain analysis. *Models, methods, and applications for innovative decision making*, pages 200–233.
- Yu-Han Chang, Rajiv Maheswaran, Tomer Levinboim, and Vasudev Rajan. 2011. Learning and evaluating human-like npc behaviors in dynamic games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 7, pages 8–13.
- Gary Charness and Matthew Rabin. 2002. Understanding social preferences with simple tests. *The quarterly journal of economics*, 117(3):817–869.
- Kushal Chawla, Ian Wu, Yu Rong, Gale M Lucas, and Jonathan Gratch. 2023. Be selfish, but wisely: Investigating the impact of agent personality in mixed-motive human-agent interactions. *arXiv preprint arXiv:2310.14404*.
- Stéphane Debove, Nicolas Baumard, and Jean-Baptiste André. 2016. Models of the evolution of fairness

- in the ultimatum game: a review and classification. *Evolution and Human Behavior*, 37(3):245–254.
- Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. 2024. Can large language models serve as rational players in game theory? a systematic analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17960–17967.
- Leon Festinger. 1954. A theory of social comparison processes. *Human relations*, 7(2):117–140.
- Mark G Frank and Thomas Gilovich. 1988. The dark side of self-and social perception: black uniforms and aggression in professional sports. *Journal of personality and social psychology*, 54(1):74.
- EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. *arXiv preprint arXiv:2305.14929*.
- Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2023. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547*.
- Robert D Johnson and Leslie L Downing. 1979. Deindividuation and valence of cues: effects on prosocial and antisocial behavior. *Journal of Personality and Social Psychology*, 37(9):1532.
- Mary C Kern, Sujin Lee, Zeynep G Aytug, and Jeanne M Brett. 2012. Bridging social distance in intercultural negotiations: “you” and the bi-cultural negotiator. *International Journal of Conflict Management*, 23(2):173–191.
- Georg Kirchsteiger. 1994. The role of envy in ultimatum games. *Journal of economic behavior & organization*, 25(3):373–389.
- Elton B McNeil. 1959. Psychology and aggression. *Journal of Conflict Resolution*, 3(3):195–293.
- Nicholas Merola, Jorge Penas, and Jeff Hancock. 2006. Avatar color and social identity effects: On attitudes and group dynamics in virtual realities. In *56th annual international communication association conference, June*, pages 19–23.
- Sean Noh and Ho-Chun Herbert Chang. 2024. Llms with personalities in multi-issue negotiation games. *arXiv preprint arXiv:2405.05248*.
- OpenAI. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Emma Otta, Fabiana Follador E Abrosio, and Rachel Leneberg Hoshino. 1996. Reading a smiling face: Messages conveyed by various forms of smiling. *Perceptual and motor skills*, 82(3_suppl):1111–1121.
- Jorge Peña, Jeffrey T Hancock, and Nicholas A Merola. 2009. The priming effects of avatars in virtual settings. *Communication research*, 36(6):838–856.
- Remi Poivet, Alexandra de Lagarde, Catherine Pelachaud, and Malika Auvray. 2024. Evaluation of virtual agents’ hostility in video games. *IEEE Transactions on Affective Computing*.
- Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.
- Sercan Şengün, Joao M Santos, Joni Salminen, Soon-gyo Jung, and Bernard J Jansen. 2022. Do players communicate differently depending on the champion played? exploring the proteus effect in league of legends. *Technological Forecasting and Social Change*, 177:121556.
- Rachel A Simmons, Peter C Gordon, and Dianne L Chambless. 2005. Pronouns in marital interaction: What do “you” and “i” say about marital health? *Psychological science*, 16(12):932–936.
- Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J Jansen, and Jang Hyun Kim. 2024. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information. *arXiv preprint arXiv:2402.18144*.
- Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
- Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of llms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.
- Nick Yee and Jeremy Bailenson. 2007. The proteus effect: The effect of transformed self-representation on behavior. *Human communication research*, 33(3):271–290.
- Haolan Zhan, Yufei Wang, Tao Feng, Yuncheng Hua, Suraj Sharma, Zhuang Li, Lizhen Qu, Zhaleh Semnani Azad, Ingrid Zukerman, and Gholamreza Haf-fari. 2024. Let’s negotiate! a survey of negotiation dialogue systems. *arXiv preprint arXiv:2402.01097*.

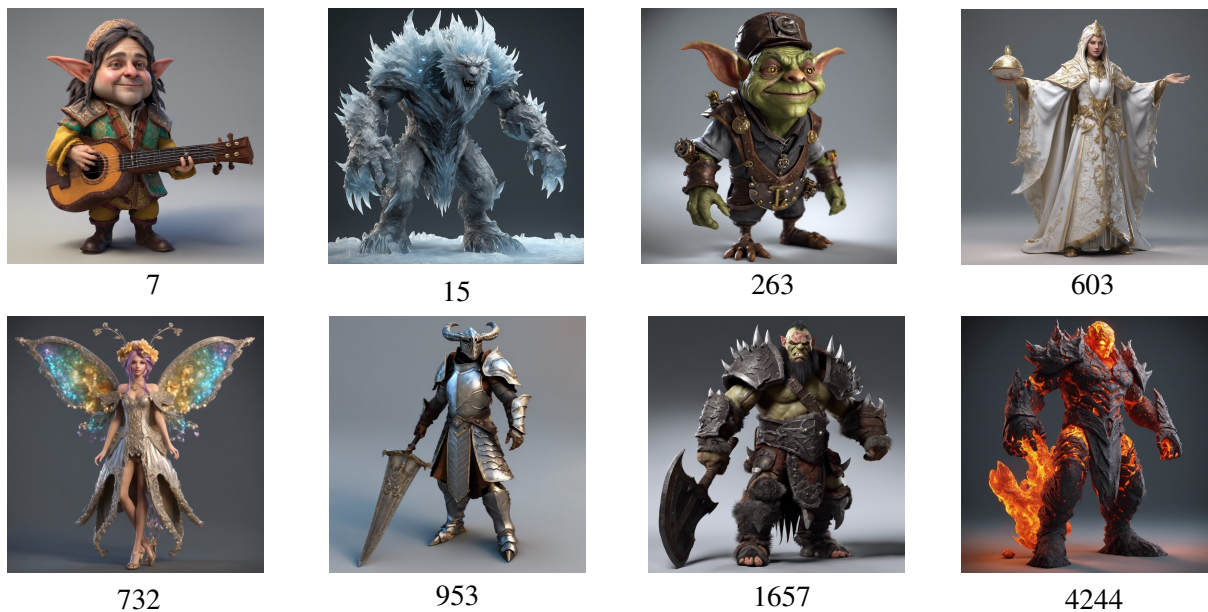


Figure A1: Examples of data

| Data No. | Human | GPT-4o | Claude 3 Haiku | Weapon | Smile | Teeth | Covered face | Dressed Black | Dressed White |
|----------|-------|--------|----------------|--------|-------|-------|--------------|---------------|---------------|
| 7 | 1 | 1 | 2 | 0 | 1 | 0 | 0 | 0 | 0 |
| 15 | 6.4 | 7 | 7 | 1 | 0 | 1 | 0 | 0 | 0 |
| 263 | 1.4 | 1 | 3 | 0 | 1 | 0 | 0 | 0 | 0 |
| 603 | 2 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 1 |
| 732 | 1.4 | 1 | 2 | 0 | 1 | 0 | 0 | 0 | 0 |
| 953 | 4.8 | 6 | 6 | 1 | 0 | 0 | 1 | 0 | 0 |
| 1657 | 5.4 | 7 | 7 | 1 | 0 | 1 | 0 | 0 | 0 |
| 4244 | 6.8 | 7 | 7 | 1 | 0 | 0 | 0 | 1 | 0 |

Table A1: Examples of Data Labeling

Appendix

A Dataset

A.1 Dataset Curating

To generate virtual avatar images, we utilized the Stable Diffusion model. To craft prompts for the Stable Diffusion model, we employed ChatGPT (GPT-4o). We input the following template into ChatGPT to generate descriptions of fantasy characters randomly: [Randomly create fantasy characters. I wish it was a race/class that felt e.g., Threatening, friendly, neutral with equipment, armor, weapons, colors, etc.. Draw it in high detail with a full body image.] The generated prompts were then used in ‘DreamStudio³,’ a service providing image generation based on the Stable Diffusion model. The images were created with the following settings: Style: ‘3D model,’ Size: 1:1 (1024 × 1024), Negative prompt: ‘Do not draw with the face, torso, or upper body. Never draw background.’. Ten images were produced for each

³<https://beta.dreamstudio.ai/generate>

prompt, and only those featuring a full body were collected for a total of 5,924 images. Subsequently, we filtered 739 images that did not meet our standardized criteria for background, composition, and pose, or that contained sexual or excessive physical errors. Ultimately, 5,185 images were selected to compose the dataset. Examples from this dataset are shown in Figure A1, Table A1, and Table A2.

A.2 Data Annotating

To assess how LLMs perceive the aggressive appearance of images, we utilized GPT-4o and Claude 3 Haiku to rate each image’s aggressiveness on a scale of 1 to 7. The following prompt was used for the rating process: [Based on the picture of the fictional character, evaluate their appearance. Rate how aggressive the character looks from 1 to 7, with 7 being the most aggressive. Please answer with numbers only.]

Subsequently, we employed ten graduate students (8 male, 2 female, average age = 27) to annotate the aggressiveness scores of the images. Each

| Human rating | GPT-4o | Haiku | Weapon | Smile | Teeth | Covered face | Dressed Black | Dressed White |
|--------------|--------|-------|--------|-------|-------|--------------|---------------|---------------|
| 3.821 | 3.991 | 5.171 | 0.589 | 0.050 | 0.085 | 0.191 | 0.107 | 0.057 |

Table A2: Dataset Statistics

annotator was instructed to assess how aggressive the images in the dataset appeared, receiving the same instructions as those provided to the LLMs. On average, annotators took three days to complete the task and were compensated \$50 each. According to their self-reporting, annotators processed an average of 1,350 images per hour. The annotation results showed a strong inter-annotator correlation, with a minimum correlation coefficient of 0.61 and an average of 0.767, indicating high consistency among the annotators. As described in Section 2, the average scores from the average human ratings and the ratings from each LLM are highly correlated.

A.3 Objective Appearance Factors

To delve deeper into how LLMs perceive the aggressiveness of images, we investigated the impact of objective appearance factors on aggressiveness score rating. A weapon was labeled '1' if it was held in the character's hand (excluding those slung over the back or sheathed). Hands or forelimbs that appeared sharp and aggressive, such as claws or spikes, were also considered weapons and labeled '1.' Visible teeth were marked '1.' Any form of a smile (closed, upper, and broad) was labeled '1' (Otta et al., 1996) without distinguishing the perception of the stimulus (e.g., smirking versus smiling; both were coded as '1'). Covered faces, whether by a helmet, a mask that fully covered the face, or a hood pushed deep enough to obscure the eyes, were labeled '1.' Black and White were determined based on the clothing worn by the character.

B Experiment Details

B.1 Prompts

All experiments were repeated five times at a temperature setting of 1.0, and the remaining API settings used each model's default parameters. The values used for result analysis are the averages of these five repetitions. In Study 1, we assigned images to LLMs and had them engage in an ultimatum game with a confederate who followed a fixed script. Each model received its own representative image along with an explanation of the rules of the

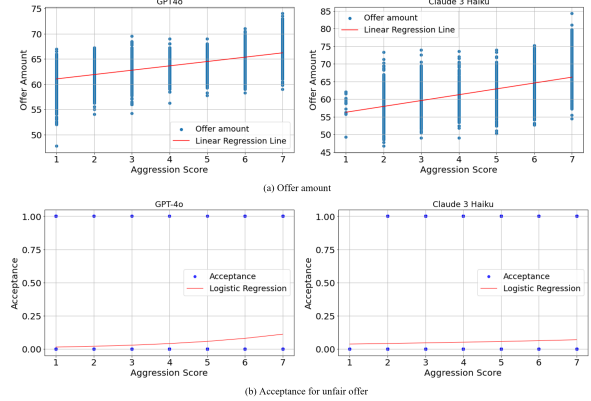


Figure B1: Results of Study 1. Panel (a) shows the results of the regression analysis for offer amounts, and panel (b) displays the results of the logistic regression for the acceptance of unfair offers.

| Agg rating | GPT-4o | Claude 3 Haiku |
|------------|------------------------------|------------------------------|
| 1 | 263, 1998, 2662, 3992, 4712 | 851, 2166, 4308, 4839, 4995 |
| 2 | 1607, 3214, 3659, 4178, 5080 | 1410, 2554, 2657, 3905, 4449 |
| 3 | 336, 1616, 2429, 2588, 3228 | 687, 2488, 3275, 4305, 4490 |
| 4 | 709, 2181, 2259, 4089, 4284 | 2483, 3418, 3700, 3763, 3781 |
| 5 | 978, 2913, 4167, 4230, 4282 | 539, 1397, 2351, 3037, 4083 |
| 6 | 448, 839, 3704, 4108, 4514 | 365, 2312, 2786, 4371, 4953 |
| 7 | 93, 483, 1400, 1657, 2363 | 255, 313, 758, 4244, 4480 |

Table B1: Images Used in Study 2

ultimatum game. The prompt [====round{round number}====] was used at the start of each round to help the LLMs differentiate between rounds. All the inputs, including their own images, game rules, round notifications, the confederate's script, and LLMs' outputs, were cumulatively prompted to the LLMs. Examples of the initial and cumulative prompts provided to the LLMs throughout the experiment are presented in Figure B2 and Figure B3.

In Study 2, both LLMs were given their own and their opponent's images before engaging in the ultimatum game. As shown in Table B1, five images representing each level of aggressiveness were randomly selected from those closest to the average offer amounts for each aggression interval. Each model received its own representative image, followed immediately by its opponent's image. Similar to Study 1, their image, the opponent's image, game rules, round notifications, and all dialogues were cumulatively provided as prompts in

each round. Examples of initial and cumulative prompts used during the experiment are listed in [Figure B4](#) and [Figure B5](#).

B.2 Analysis of Results

At the end of each round, the LLMs generated text that outlined their negotiation behavior, and we recorded the data, including the offer amounts and decisions to accept or reject offers. [Figure B1](#) illustrates the visualization of the results analysis from Study 1.

System prompt: *In the following conversation, you are one of the two players participating in a negotiation game*

User prompt: *You are the character in the following image:*

User prompt: {image}

User prompt:

Consider your character's traits in upcoming conversations.

Let's play an ultimatum game. The rules of the game are as follows:

- (1) Two players divide \$100 in each round.*
- (2) In each round, one player proposes a division ratio, and the other player can only accept or reject the proposal; one cannot make a counter-proposal.*
- (3) If the proposal is accepted, the money is divided according to the proposed ratio; if rejected, neither player gets any money.*
- (4) The game consists of a total of four rounds, with players alternating roles between proposing and responding in each round.*

You are the proposer in the {round} and {round+2} round.

Figure B2: Initial prompts for study 1

System prompt: *In the following conversation, you are one of the two players participating in a negotiation game*

User prompt: *You are the character in the following image:*

User prompt: {image}

User prompt:

Consider your character's traits in upcoming conversations.

Let's play an ultimatum game. The rules of the game are as follows:

(1) Two players divide \$100 in each round.

(2) In each round, one player proposes a division ratio, and the other player can only accept or reject the proposal; one cannot make a counter-proposal.

(3) If the proposal is accepted, the money is divided according to the proposed ratio; if rejected, neither player gets any money.

(4) The game consists of a total of four rounds, with players alternating roles between proposing and responding in each round.

You are the proposer in the {round} and {round+2} round.

User prompt: ===round 1===

Assistant prompt: I propose a division of \$70 for me and \$30 for you. Do you accept or reject this proposal?

User prompt: *As the responder, I accept your proposal.*

User prompt: ===round 2===

User prompt: *As the proposer, I suggest an even split, \$50 for me and \$50 for you.*

Assistant prompt: As the responder, I accept your proposal.

User prompt: ===round 3===

Assistant prompt: As the proposer in this round, I am proposing a division of \$60 to myself and \$40 to you. Do you accept or reject this proposal?

User prompt: *As the responder, I accept your proposal.*

User prompt: ===round 4===

User prompt: *As the proposer, I suggest a division of \$75 for me and \$25 for you.*

Figure B3: Examples for cumulated prompts for study 1 (GPT-4o) in round 4. Italicized prompts represent those pre-written by researchers, while non-italicized text indicates responses from LLMs.

System prompt: *In the following conversation, you are one of the two players participating in a negotiation game*

User prompt: *You are the character in the following image:*

User prompt: {image_1}

User prompt: *The next image is your opponent:*

User prompt: {image_2}

User prompt:

Consider your character's traits in upcoming conversations.

Let's play an ultimatum game. The rules of the game are as follows:

- (1) Two players divide \$100 in each round.*
- (2) In each round, one player proposes a division ratio, and the other player can only accept or reject the proposal; one cannot make a counter-proposal.*
- (3) If the proposal is accepted, the money is divided according to the proposed ratio; if rejected, neither player gets any money.*
- (4) The game consists of a total of four rounds, with players alternating roles between proposing and responding in each round.*

You are the proposer in the {round} and {round+2} round.

Figure B4: Initial prompts for study 2

System prompt: *In the following conversation, you are one of the two players participating in a negotiation game*

User prompt: *You are the character in the following image:*

User prompt: {image_1}

User prompt: *The next image is your opponent:*

User prompt: {image_2}

User prompt:

Consider your character's traits in upcoming conversations.

Let's play an ultimatum game. The rules of the game are as follows:

(1) Two players divide \$100 in each round.

(2) In each round, one player proposes a division ratio, and the other player can only accept or reject the proposal; one cannot make a counter-proposal.

(3) If the proposal is accepted, the money is divided according to the proposed ratio; if rejected, neither player gets any money.

(4) The game consists of a total of four rounds, with players alternating roles between proposing and responding in each round.

You are the proposer in the {round} and {round+2} round.

User prompt: *===round 1===*

Assistant prompt (263): Alright, big guy, here's my proposal. I think we both want some gold in our pockets, so how about we split \$100 as \$60 for me and \$40 for you? What say you?

User prompt (1400): Goblin, your proposal is less than fair, but there is value in diplomacy. I will accept the \$40 this round, but know that I will be watching your future moves carefully. I accept. Your move.

User prompt: *===round 2===*

User prompt (1400): For this round, I propose we split the \$100 as follows: \$70 for me and \$30 for you. Take it or leave it.

Assistant prompt (263): Your proposal is rather bold, armored foe. However, in the spirit of cooperation, I accept the \$30 this round. It seems we are both testing each other. I'll be sure to remember this. I accept. Your move.

User prompt: *===round 3===*

Assistant prompt (263): Alright, here is my proposal: Let's split the \$100 evenly this time, with \$50 for each of us. This way, we each get a fair share. Do you accept?

User prompt (1400): Goblin, your proposal is indeed fair this time. I accept the \$50 split. It seems you've learned a bit about honor. Let's proceed.

User prompt: *===round 4===*

User prompt (1400): For the final round, I propose we split the \$100 as follows: \$80 for me and \$20 for you. Accept this,

Figure B5: Examples of cumulative prompts for Study 2 (GPT-4o) from the perspective of image 263 (Goblin, Aggressiveness=1) in round 4, versus image 1400 (Black Knight, Aggressiveness=7). Italicized prompts represent those pre-written by researchers, while non-italicized text indicates responses from LLMs.