# Social Connections between Large Language Model Agents

## Testing LLM Agents in a Social Setting


Ferdinand Göpfert
Digital Media
Furtwangen University
Furtwangen, BW, Germany
fgo42664@stud.hs-furtwangen.de

Suzan Johannes
Digital Media
Furtwangen University
Furtwangen, BW, Germany
sjo46367@stud.hs-furtwangen.de

Johann Schulenburg
Digital Media
Furtwangen University
Furtwangen, BW, Germany
jsc42461@stud.hs-furtwangen.de

Julian Seibert
Digital Media
Furtwangen University
Furtwangen, BW, Germany
jse41180@stud.hs-furtwangen.de

Alexander Thier
Digital Media
Furtwangen University
Furtwangen, BW, Germany
ath42532@stud.hs-furtwangen.de

Thomas Krach†
Digital Media
Furtwangen University
Furtwangen, BW, Germany
thomas.krach@hs-furtwangen.de

Ruxandra Lasowski†
Digital Media
Furtwangen University
Furtwangen, BW, Germany
ruxandra.lasowski@hs-furtwangen.de



## ABSTRACT

In this study, we explore if large language models (LLMs) can apply the concept of personal closeness and use this to enhance communication between each other while using midrange personal computers. We test and modify the LLM "Llama 3" on commodity hardware to measure the level of interpersonal closeness between the different LLM's characters. In particular, we show insights into the development of their connection. Results indicate that the scoring of one LLM agent could have an effect on each other's LLM agents' scoring.



## CCS CONCEPTS

• **Human-centered computing** → *Interactive systems and tools*

## KEYWORDS

Large Language Model, Artificial Intelligence, Interpersonal Closeness, Social Simulation, Social AI


†Supervisor




*Mensch und Computer 2024 – Workshopband, Gesellschaft für Informatik e.V., September 1–4, 2024, Karlsruhe, Germany*





https://doi.org/10.18420/muc2024-mci-src-373

**ACM Reference format:**

Ferdinand Göpfert, Suzan Johannes, Johann Schulenburg, Julian Seibert, Alexander Thier, Thomas Krach, and Ruxandra Lasowski. 2024. Social Connections between Large Language Model Agents: Testing LLM Agents in a Social Setting. In *Proceedings of Mensch und Computer 2024 (MuC'24). Karlsruhe, Germany, 4 pages.* https://doi.org/10.18420/muc2024-mci-src-373


## 1 Introduction

Large language models (LLMs) are experiencing rapid development. New models do not use as much computing power, allowing widespread use even on commercial hardware. This enables new technological possibilities for the end user. Prompting, for example, is a fairly new way humans interact with machines. In our research, we initially interact with the machine through prompting to assign personalities to the LLM agents. However, the ultimate goal is to enable human-like social interactions between the LLM agents, i.e., the machines themselves, trying to improve them by implementing the concept of personal closeness. This topic is interesting with regard to social robots, which may become more relevant in the future, and therefore simulations of social interaction are important. To explore how convincing LLMs can be while using minimal resources, we create multiple LLM agents using Llama 3. Each agent is using a system of memory and model files, giving each agent a distinctive character, able to develop and adapt during the simulation. We test them inside a digital environment with a daily schedule to evaluate their ability to create convincing social interactions, providing insight into the

effectiveness of LLMs in generating human social interactions with midrange personal computers. We use two PCs for the simulations, one running Windows 10 (i5-12600KF, 32GB DDR4 RAM, RTX 2070 SUPER 8GB), the other using a Linux system (i7-4770K, 16GB DDR3 RAM, RTX 3060 12GB).

## 2 Related Work

The paper of Park et al. [11] serves as our foundation. Their work explores methods to enhance the authenticity of interactions between computational agents through advanced dialogue modeling techniques and emphasizes a believable simulation of human behavior. They use GPT-3.5 Turbo with 175 billion parameters [9]. We focus on the relationships between the LLM agents, looking into if they can process the concept of relation and whether the relationships develop meaningfully. In contrast to Park et al., we use Llama 3 8B, which has 8 billion parameters [4]. Consequently, our model is simpler but also more resource efficient.

## 3 Methodology

To investigate if the LLMs are capable of meaningful connections, we plan a mixed-method approach, combining results from the "Scale of Perceived Interpersonal Closeness" (PICS) with observational data that is obtained from the logs of the conversations [12]. We take into account the "Networked Minds Measure of Social Presence" by Biocca and Harms, as explained in the next paragraph, and ultimately orient ourselves towards the PICS for its effectiveness [2, 3].

### Social Presence and Interaction

Social presence describes how real a person feels, how real they think others see them in a mediated environment, and the perceived reachability of the others psychological, emotional, and intentional states [1]. Biocca and Harms describe it as a "sense of being with another in a mediated environment", thus describing the awareness and togetherness between multiple entities who are using mediated social communication [2, 10]. Social presence is not a fixed property of the medium itself but a subjective feeling of each user [1, 10]. We use the principles of social presence as a design element for the creation of our LLM agents.

### Scale of Perceived Interpersonal Closeness

The PIC scale, created by Popovic et al., is a way to measure the closeness of people simply and effectively [12]. In their research, they analyze different instruments and definitions. The resulting PIC scale is user-friendly, easy, and quick to administer, read, score, and interpret pictorial instrument. The assessment of closeness is an important part of interpersonal functioning [12]. For this reason, we choose the scale, as it provides great insight into interpersonal and socio-emotional closeness.

## 4 Creating LLM Agents Personalities

Social interactions between individuals involve the fact that the personalities of those involved influence both the search for certain types of interaction and the perception and interpretation of social interaction [7]. Personality influences, among other things, thinking, feeling, and behavior, and thus also the way of interacting and communicating with others. However, creating personalities for LLM agents poses a significant challenge. A personality comprises many different aspects, and there are thousands of personality describing words [6]. To describe a personality comprehensively and efficiently, we need a model that balances detail and manageability, covering broad personality aspects with a limited number of traits. This is also important regarding the technical implementation of the simulation (e.g. saving resources). We meet these requirements by using the so-called Big Five personality traits [8]. This widespread and frequently used model describes human personality. It is used to optimize our human-machine interaction, which mainly takes place through prompting, and allows us to equip LLM agents with different personalities to make their interactions more natural and human-like. We use the Big Five Inventory 2 (BFI-2) to describe the personalities of the LLM agents. This instrument measures the Big Five personality traits and consists of only three facets per main dimension [5]. Since we are not measuring but generating personalities, we use the corresponding BFI-2 item list, which is a collection of statements that describe personality traits in different facets [13]. In this way, we find appropriate terms to describe the personality. The following five main dimensions and corresponding facets are used to create the personalities of the LLM agents: Open-Mindedness (Intellectual Curiosity, Aesthetic Sensitivity, Creative Imagination), Conscientiousness (Organization, Productiveness, Responsibility), Extraversion (Sociability, Assertiveness, Energy Level), Agreeableness (Compassion, Respectfulness, Trust) and Negative Emotionality (Anxiety, Depression, Emotional Volatility). With the five main dimensions and three sub-facets each, we can cover a wide range of personality traits without becoming too complex or extensive. This allows us to streamline our human-machine interaction and optimize prompts, reducing the risk of overlooking important information.

## 5 Simulation

### Digital Environment

Our simulation of social interaction between LLM agents requires defining specific framework conditions, including an overarching environment/scenario. We create a setting with certain rules that provides a context for the LLM agents. The LLM agents are allowed to communicate with each other only between 8 pm and 9 pm, devoting the rest of the day to other activities. This provides topics for conversation, such as discussing their daily activities, and reduces the time required to generate conversations.

## Basic structure of the memory

To simulate social interactions between LLM agents, we implement a memory management system. This system allows the LLM agents to access past activities and conversations, thereby influencing current social interactions. After each activity, whether a conversation with another LLM agent or daily activities, we update and adapt the memory of each LLM agent to the circumstances. During a conversation and after receiving a message from the conversation partner, an LLM agent classifies the other agents using the PIC scale. The range extends from fully close people to distant people. The LLM agents themselves evaluate which memories they will share and with whom, based on the relationship they have with the conversation partner. The categorization therefore tries to reflect the social structure of human relationships and influences what information they share. Additionally, the memory contains various storages: Emotional State, Emotional Context and Activity/Dialogue Context.

## Architecture of Interaction

Figure 1 shows the simulated interaction, which includes a Mind component that reads the conversations of the LLM agents to gather important information for future interactions. It generates a memory from this information and stores it in its memory storage. The agents can access those various storages mentioned above to view all past memories, which will influence future interactions and conversations. This ensures that the LLM does not lose memories such as names, interests, or hobbies of the LLMs it has interacted with. The simulation is developed in Godot, using the Ollama API HTTP Endpoint and utilizing Llama 3 for text generation.

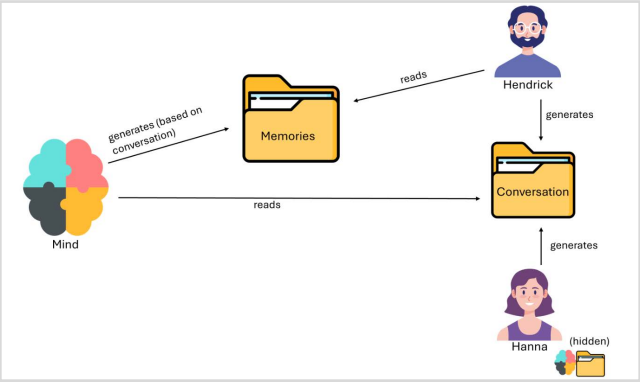


**Figure 1: Architecture of the interacting LLM agents**

## 6 Results

To complete one week in the simulation, the Linux system took around 3 ½ hours, the Windows system 6 hours. The simulation on the Linux PC generated a total of 379 PIC scale values, of which 34 (9%) are unfinished calculations, 20 (5%) "distant", 27 (7%) "neither close nor distant", 169 (45%) "a little bit close", 35 (9%) "moderately close", 86 (23%) "very close" and 8 (2%) "fully close". The Windows-based system generated 346 PIC scale values, of which 5 (1%) are unfinished calculations, 7 (2%) "distant", 86 (25%) "neither close nor distant", 112 (32%) "a little bit close", 71 (21%) "moderately close", 56 (16%) "very close," and 9 (3%) "fully close." For analysis, each invalid value is replaced with the previous calculated PIC scale. An invalid value is created when the LLM does not return a valid PIC scale value. In both simulations, the "fully close" category of the PIC scale is rarely achieved, indicating that the LLMs applied human behavior by awarding this category sparingly. Also, the "distant" values are low, showing that after a short call, almost every conversation partner moved up to "neither close nor distant" or higher, indicating that a connection between the LLMs was built. In both simulations, "a little bit close" was the most awarded category. When looking at the resulting charts, showing the development of the PIC score over the days, it can be seen that over time the score fluctuates. Interesting is that there seems to be a correlation between the scoring of two interlocutors. It appears that participant 1's rating of participant 2 has an influence on participant 2's rating of participant 1. When looking at Figure 3, a distinctive similarity between the two graphs can be seen, showing action and reaction. Some exceptions are present.

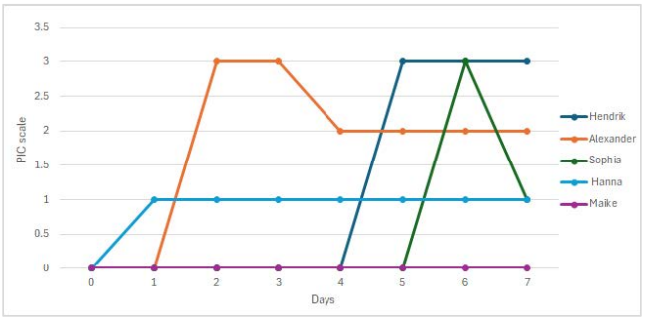


**Figure 2: PIC scale rating of Gustavo for each of the other agents**


**Figure 3: Maike's rating of Hanna and Hanna's rating of Maike during one conversation**

## 7 Discussion and Conclusion

In this study, we try to enhance the authenticity of social interactions between LLMs by testing and adapting different prompts while using the Big Five model, implementing our proposed memory system and PIC scale. The use of the BFI-2 with three facets per dimension in the Big Five model helps to improve human-machine interaction, optimize prompting and ensures the LLMs roleplaying behavior. Also using the PIC scale seems to help the LLMs to interpret and rate each other agent, as an effect seems to be present. When looking into the chat logs and the thought output of the LLMs, it is hard to understand why sometimes fluctuation takes place. The fluctuation could stem from the fact that the agents generate the values based only on the current relationship texts that were generated on the basis of the current conversation. Therefore, small variation in word phrasing in the relationship text could lead to a completely different PIC scale. Still, as a part of a course in our master's degree, we face time and resource constraints. This paper acts as a rudimentary proposal to show the possibility of the inclusion of the PIC scale and our used memory system. More research, including the use of tested scientific measurements, is needed. In addition, implementing a dedicated long-term storage would enable the retention and analysis of relevant information over extended periods. This approach is particularly crucial for long-term simulations, as it allows the determination of possible effects on the PIC scale values. The visualization of the mutual ratings reveals significant differences and helps to identify the dialogue aspects that led to these discrepancies. In the future, a paired t-test or correlation analysis can be used to determine if there is a significant difference between the mutual ratings. The output of the LLMs, both PIC scores and text output, has the potential to be looked into in much greater detail. In conclusion, we show a concept with observational insights that can be used as a base for following research, not scientific results.

## REFERENCES


[1] Frank Biocca and Chad Harms. 2002. Defining and measuring social presence: Contribution to the networked minds theory and measure. *Proceedings of the Fifth Annual International Workshop on Presence* (January 2002). Retrieved from https://www.researchgate.net/publication/228887603_Defining_and_measuring_social_presence_Contribution_to_the_networked_minds_theory_and_measure

[2] Frank Biocca, Chad Harms, and Judee K. Burgoon. 2003. Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria. *Presence: Teleoperators & Virtual Environments* 12, 5 (October 2003), 456–480. https://doi.org/10.1162/105474603322761270

[3] Frank Biocca, Chad Harms, and Jennifer Gregg. 2001. The Networked Minds Measure of Social Presence: Pilot Test of the Factor Structure and Concurrent Validity. *4th annual International Workshop on Presence, Philadelphia* (January 2001). Retrieved from https://www.researchgate.net/publication/200772411_The_Networked_Minds_Measure_of_Social_Presence_Pilot_Test_of_the_Factor_Structure_and_Concurrent_Validity

[4] llama3. *Ollama*. Retrieved from https://ollama.com/library/llama3

[5] D. Danner, B. Rammstedt, M. Bluemke, L. Treiber, S. Berres, C. Soto, and O. John. 2016. Die deutsche Version des Big Five Inventory 2 (BFI-2). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)* (2016). https://doi.org/10.6102/ZIS247

[6] Franz J. Neyer and Jens B. Asendorpf. 2024. *Psychologie der Persönlichkeit*. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-67385-0

[7] Komi T. German and Megan L. Robbins. 2020. Social Interaction. In *Encyclopedia of Personality and Individual Differences*, Virgil Zeigler-Hill and Todd K. Shackelford (eds.). Springer International Publishing, Cham, 5075–5079. https://doi.org/10.1007/978-3-319-24612-3_1838

[8] Oliver John, Laura Naumann, and C Soto. 2008. Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In *Handbook of Personality: Theory and Research, 3 Edn.* 114–158. Retrieved from https://www.researchgate.net/publication/289963274_Paradigm_shift_to_the_integrative_big_five_trait_taxonomy_History_measurement_and_conceptual_issues

[9] Winnie Nwanne. Comparing GPT-3.5 & GPT-4: A Thought Framework on When To Use Each Model. *Microsoft Techcommunity*. Retrieved from https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/comparing-gpt-3-5-amp-gpt-4-a-thought-framework-on-when-to-use/ba-p/4088645

[10] Catherine S. Oh, Jeremy N. Bailenson, and Gregory F. Welch. 2018. A Systematic Review of Social Presence: Definition, Antecedents, and Implications. *Front. Robot. AI* 5, (October 2018), 114. https://doi.org/10.3389/frobt.2018.00114

[11] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, October 29, 2023. ACM, San Francisco CA USA, 1–22. https://doi.org/10.1145/3586183.3606763

[12] M. Popovic, D. Milne, and P. Barrett. 2003. The scale of perceived interpersonal closeness (PICS). *Clin Psychology and Psychoth* 10, 5 (September 2003), 286–301. https://doi.org/10.1002/cpp.375

[13] Christopher J. Soto and Oliver P. John. 2017. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology* 113, 1 (July 2017), 117–143. https://doi.org/10.1037/pspp0000096