Wizard-of-Oz vs. GPT-4: A Comparative Study of Perceived Social Intelligence in HRI Brainstorming

Anita Vrins*
A.M.Vrins@tilburguniversity.edu
Tilburg University
The Netherlands

Ethel Pruss
E.Pruss@tilburguniversity.edu
Tilburg University
The Netherlands

Caterina Ceccato
C.Ceccato@tilburguniversity.edu
Tilburg University
The Netherlands

Jos Prinsen
J.M.Prinsen@tilburguniversity.edu
Tilburg University
The Netherlands

Alwin de Rooij
AlwindeRooij@tilburguniversity.edu
Tilburg University
The Netherlands

Maryam Alimardani
M.Alimardani@tilburguniversity.edu
Tilburg University
The Netherlands

Jan de Wit
J.M.S.deWit@tilburguniversity.edu
Tilburg University
The Netherlands

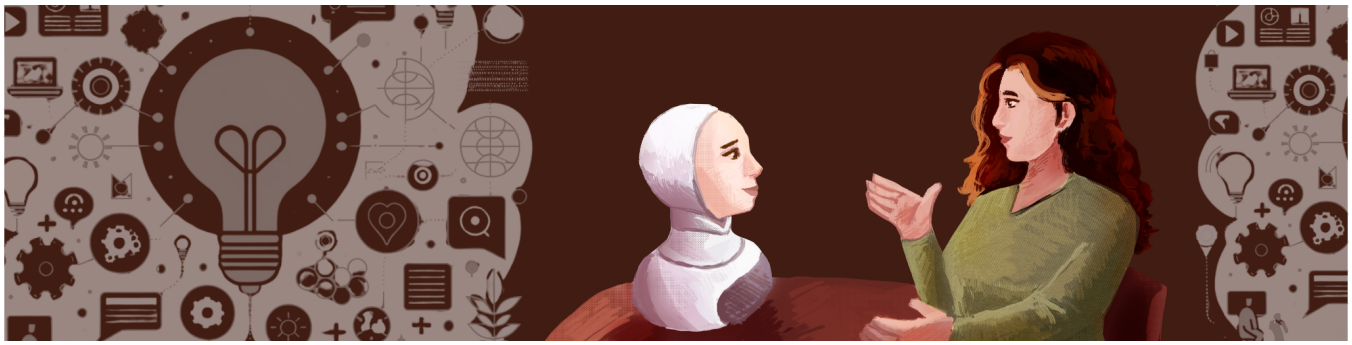

Figure 1: Illustration of the experimental setup. Participants brainstormed with a Furhat robot under two experimental conditions: 1) the robot was controlled by a Wizard, 2) the robot used a large language model (GPT-4) to generate responses.

ABSTRACT

Human-robot interaction often employs the Wizard-of-Oz (WoZ) paradigm, where a human controls the robot. However, this approach has limitations, such as a lack of autonomy that impedes real-world applications. Large language models (LLMs) can replace WoZ in conversational tasks, such as brainstorming. We propose that, in such application domains, LLM-controlled robots can achieve comparable perceived social intelligence to WoZ-controlled robots. An experiment (n=27, within-subject design) tested this by having participants brainstorm with an LLM- and WoZ-controlled Furhat robot. Bayesian analyses revealed substantial evidence for the null model for perceived social intelligence, social presentation, and social information processing, indicating similar perceptions of social intelligence for WoZ- and LLM-controlled robots. Participants tentatively preferred the LLM-controlled robot, and reliably identified when the robot was WoZ- or LLM-controlled. This study highlights the potential of LLMs to replace the WoZ paradigm and transform HRI in various research and application domains.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

Human-Robot Interaction, Wizard-of-Oz, Large Language Models, Brainstorming, Social Intelligence

*Authors AV and EP contributed equally to this research.

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0323-2/24/03
https://doi.org/10.1145/3610978.3640755

ACM Reference Format:
Anita Vrins, Ethel Pruss, Caterina Ceccato, Jos Prinsen, Alwin de Rooij, Maryam Alimardani, and Jan de Wit. 2024. Wizard-of-Oz vs. GPT-4: A Comparative Study of Perceived Social Intelligence in HRI Brainstorming. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610978.3640755

1 INTRODUCTION

The ability to have a natural conversation with an autonomous agent so far has been limited in the context of Human-Robot Interaction (HRI). The experience of a dynamic and human-like dialogue is often realized through the "Wizard-of-Oz" (WoZ) paradigm. In this paradigm, the social agent is remotely controlled by a human behind the scenes [13]. However, conversational robots controlled by WoZ often struggle to keep up with the user's speed and level of detail in speech. This paradigm constrains the usage of such agents in laboratory settings and lacks the autonomy that is crucial for real-world application.

The incorporation of Large Language Models (LLMs), such as GPT, in HRI, holds the potential to help overcome existing barriers and facilitate more effective interactions between humans and robots compared to traditional human-operated or scripted interactions [18]. LLMs have demonstrated impressive capabilities in tasks related to natural language generation and understanding, as evidenced in previous studies [3, 10]. However, despite their potential, empirical studies on the use of LLMs for HRI are scarce [14, 8].

One promising application domain for LLM-driven robots is in brainstorming tasks [7], where the robot autonomously generates responses that blend existing knowledge with novel concepts to facilitate the production of ideas and problem solutions [4]. The success of such robots hinges on the perception of the robot as a valuable partner. As outlined by [12], humans tend to evaluate and mistrust robots in the same way they would assess another human, particularly when it comes to making mistakes. Thus, for a robot to be considered a viable brainstorming partner, it needs to approach levels of social intelligence comparable to humans.

Brainstorming relies on social information processing [11], i.e. the ability to recognize, adapt to, and predict the emotions, motivations, and cognition of others [2]. To effectively build on each other's ideas, it is crucial to identify, adjust to, and anticipate the (social) implications of ideas, and interpret the social signals that accompany their expression [5, 11]. This guides how to respond in a manner that facilitates the desired direction and effectiveness of brainstorming. In turn, social presentation, i.e., the ability to present oneself (and one's ideas) in a socially desirable and intelligent manner [2], facilitates further uptake of one's ideas by others. This exemplifies the importance of social intelligence in HRI applications, such as brainstorming.

This paper explores how LLM-controlled robots, compared to WoZ-controlled robots, are perceived in collaborative brainstorming tasks concerning their social intelligence and user preference. The research questions are as follows:

RQ1: Do users perceive the LLM-controlled robot as socially intelligent as the WoZ-controlled robot?

RQ2: Which paradigm do users prefer for a collaborative brainstorming task; the LLM-controlled robot or the WoZ?

2 METHOD

2.1 LLM-Controlled Social Robot

To explore the research questions we integrated an LLM and developed a WoZ setup to control the "Furhat" robot. The Furhat is a human-like robotic head known for its back-projected facial animations and sophisticated speech recognition and synthesis capabilities [1]. Integration of the LLM was done by connecting the GPT-4 API to the Furhat API, enabling the robot to utilize GPT-4 for generating tailored responses during interactions with users.

2.1.1 Prompt Engineering. A prompt was engineered for the LLM such that the robot would brainstorm with the user on a certain topic, ensuring the robot could effectively process human ideas into prompts and then generate appropriate responses. The prompt consisted of three main parts. Firstly, the robot was informed of the brainstorming task and that it should be addressing one user. Secondly, the robot should both provide new ideas and build upon the ideas that the user presented. Thirdly, the robot should keep the conversation on the topic and steer away from engaging in conversation about personal or sensitive information.

2.1.2 Wizard of Oz. The Wizard conducted a brainstorming session with the user, following a strategy similar to the instructional prompt provided to the LLM. The objective was to have a natural conversation and expand on the user's ideas when possible. Unlike the robot, the Wizard had direct access to the user's speech through audio but was limited by manual typing speed when transmitting messages to the Furhat robot. The robot used a text-to-speech API to vocalize the transmitted messages.

2.2 Experiment

We experimentally tested whether the LLM-controlled robot was perceived to be as socially intelligent as the WoZ-controlled robot (RQ1) and identify which robot users preferred for the brainstorming task (RQ2).

2.2.1 Participants. The study was conducted with 27 university students (19 females, 7 males, 92% in the age range of 18-24 and 8% in the range of 25-34, demographic information for one participant was not available). A majority of the participants (n = 21) had no previous experience with social robots, while 5 reported having limited experience. Ethical approval for the study was granted by the Research Ethics Committee of our institution.

2.2.2 Protocol. The experiment had a within-subjects design with two randomized conditions: 1) a WoZ-controlled robot, where two experienced "Wizards" were randomly assigned to control the robot's responses, and 2) an LLM-controlled robot (GPT-4) (see Figures 2 and 3). Together with the robot, participants brainstormed solutions to "social isolation in university students" and "work-life balance for university students." Each brainstorm lasted 5 minutes and was preceded by task instructions. The conditions and topics were administered in random order.

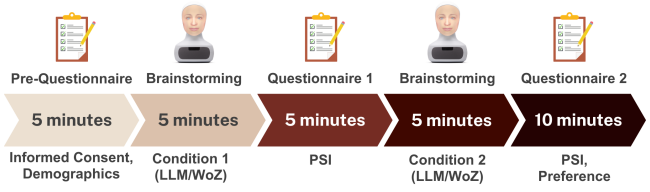

Figure 2: An overview of the experiment timeline.

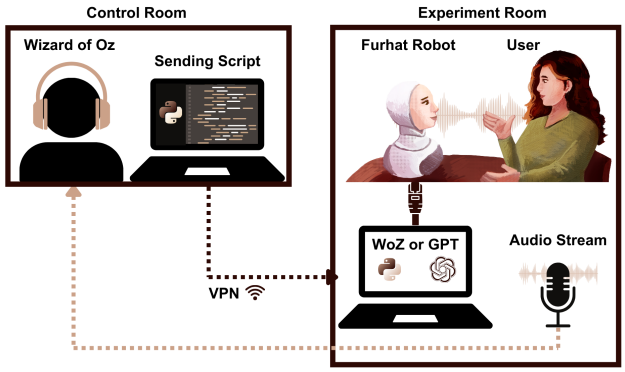

Figure 3: The experiment setup for both conditions. In the LLM condition, GPT-4 was used to generate responses based on the participant's speech, which was transcribed by Furhat's integrated speech recognition (based on the Microsoft Azure Speech to Text service), before being passed to GPT-4 as a prompt. In the WoZ condition, the Wizard in the control room listened to a live audio feed from the experiment room and responded to the participant by typing prompts, which were then vocalized by Furhat.

2.2.3 Instruments. As seen in Figure 2, after each task, participants filled in the Perceived Social Intelligence (PSI) scale [2]. The PSI scale captures 17 sub-constructs of social intelligence, each of which is assessed by rating 4 items on a 5-point Likert scale. These were aggregated into the three main constructs under study: overall perceived social intelligence, social presentation, and social information processing. At the end of the experiment, participants indicated their preference between the two conditions, and after being informed about the nature of the conditions, we assessed their ability to identify which robot was WoZ- or LLM-controlled. The latter was assessed to explore whether participants could identify differentiating features between the conditions that could, e.g., have a bearing on user preference.

2.3 Data Analysis

Bayesian statistics were used to analyze the data enabling quantifying uncertainty about inferences based on probabilities (Bayes factor) [16]. This is a more robust and flexible approach than the traditionally used Frequentist statistics for testing equivalence (RQ1, seeking evidence for the null model) and difference (RQ2, seeking evidence for the alternative model). A Bayes factor (BF_{10}) of 1 signifies no evidence for any conclusion, approaching 0 supports the null model (equivalence), and values greater than 1 increasingly indicate evidence for the alternative model (meaningful difference).

Bayesian repeated measures ANOVAs were used to analyze the PSI scores and Bayesian Multinomial and Binomial tests were used to assess the participants' preferences and their ability to identify the experimental conditions. Default priors were used. The assumption checks suggested no assumption violations. We follow [9] for interpreting the strength of the evidence indicated by BF_{10}.

3 RESULTS

3.1 Perceived Social Intelligence

The comparative analysis of the robot's perceived social intelligence showed only marginal differences between the experimental conditions (see Table 1). For Social Information Processing, Social Presentation, and Overall Perceived Social Intelligence, mean scores were slightly higher in the WoZ condition (3.163, 2.909, and 3.058, respectively) compared to the LLM condition (3.138, 2.884, and 3.033, respectively). The Bayesian repeated measures ANOVAs yielded Bayes factors (BF_{10}) of 0.287, 0.320, and 0.310 for each construct respectively, all suggesting substantial evidence for the null model. These findings indicate that participants perceived the WoZ- and LLM-controlled robot to be similarly socially intelligent.

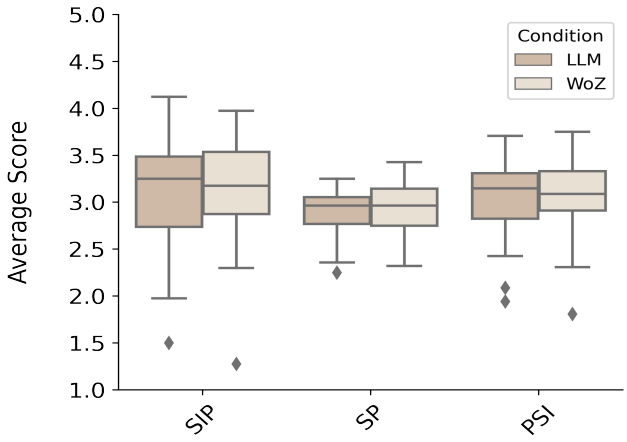

Figure 4: Boxplot of the Perceived Social Intelligence Scales: SIP = Social Information Processing, SP = Social Presentation, PSI = Overall Perceived Social Intelligence.

3.2 Preference

The Bayesian Multinomial test, conducted to assess participant preferences, yielded a Bayes Factor (BF_{10}) of 2.183 (see Table 2), suggesting anecdotal evidence for the alternative model, in favor of a preference for interacting with the LLM-controlled robot. Fifteen participants preferred the LLM-controlled robot, 8 the WoZ-controlled robot, and 4 expressed no preference. These responses were collected while participants were not yet informed about the nature of the conditions.

3.3 Ability to Identify Conditions

After being informed of the experimental conditions, the participants were asked to identify when they believed they interacted with the WoZ- and LLM-controlled robot. The Bayesian Binomial Test yielded a Bayes Factor (BF_{10}) of 5.398 (see Table 3). This indicated substantial evidence for the alternative model in favor of the ability to identify the conditions. Twenty out of 27 participants correctly identified which robot was LLM- and WoZ-controlled.

Table 1: Descriptive Statistics and Bayes Factors for Social Information Processing, Social Presentation, and Overall Perceived Social Intelligence

Scale	Conditions	N	Mean	SD	SE	95% Credible Interval		BF_{10}
						Lower	Upper	
Social Information Processing	LLM	27	3.138	0.611	0.118	2.896	3.380	0.287
	WoZ	27	3.163	0.594	0.114	2.928	3.398	
Social Presentation	LLM	27	2.884	0.255	0.049	2.783	2.985	0.320
	WoZ	27	2.909	0.278	0.053	2.799	3.019	
Overall Perceived Social Intelligence	LLM	27	3.033	0.428	0.082	2.864	3.202	0.310
	WoZ	27	3.058	0.421	0.081	2.892	3.225	

Table 2: Bayesian Multinomial Test for Preference

Level	Counts	Proportion	BF_{10}
LLM	15	0.556	
No preference	4	0.148	2.183
WoZ	8	0.296	

Table 3: Bayesian Binomial Test for Condition Identification

Level	Counts	Proportion	BF_{10}
Correct	20	0.741	5.398
Incorrect	7	0.259	

4 DISCUSSION

Our study presents groundbreaking findings on integrating LLMs in human-robot interaction, particularly in the domain of perceived social intelligence. The results indicate substantial evidence that participants perceived the LLM-controlled robot on par with the WoZ-controlled robot in terms of their social intelligence. This highlights the potential of LLMs for replacing the WoZ paradigm in tasks relying on conversational interactions, such as brainstorming; and represents a step toward developing truly autonomous socially intelligent robots in both research and real-world applications [18].

The results showed anecdotal evidence favoring a preference for the LLM-controlled robot, partly due to the LLM's superior speed and ability to offer more ideas, context, and encouragement within the same time frame. To enhance fairness in comparing a WoZ- to LLM-controlled robot, a standardized delay time for responses may be considered in future research. Note, however, that the LLMs' speed and fluency are also key advantages in HRI [17]. The preliminary results indicate a need for more extensive research involving a larger sample size to verify the impact [9]. Should future studies validate this trend, it would further highlight the potential of LLMs to enhance HRI, possibly improving its richness and vitality.

Interestingly, the results also showed that most participants correctly identified when the robot was LLM- or WoZ-controlled. Given the anecdotal evidence for an effect on user preference, it can be inferred that certain human-like characteristics distinguish WoZ- and LLM-controlled robots, such as slower and shorter responses, but may not necessarily favorably shape user preferences.

This study's insights are important for future HRI developments. The ability of LLMs to endow robots with a level of social intelligence that is perceived as identical to a human operator can significantly change the landscape of HRI applications. We speculate that the implications of this study reach beyond brainstorming tasks, and likely translate to other application domains that rely strongly on conversational interactions, such as education, healthcare, and customer service [6, 15]. Our findings suggest a future where LLM-controlled robots can autonomously engage in meaningful, dynamic, and productive interactions with humans.

The research of [14] aligns with ours, suggesting a future with less reliance on the WoZ method. In addition to the various HRI applications of GPT proposed by [14], our work extends to a comprehensive brainstorming interaction using GPT-4. This latest iteration of GPT shows a marked improvement in response quality, addressing the issues of earlier versions such as GPT-3.5, which [8] highlighted. Our findings indicate a reduction in delayed, superficial, or erroneous responses, although we note this was within the confines of a more restricted scope regarding time and topic.

However, similar to [8], we encountered issues with response timing, which are likely attributable to the Furhat robot's limitations, rather than the LLM. The robot's reliance on cloud-based services for speech recognition and synthesis can also affect performance due to possible service disruptions and transcription inaccuracies. Regarding limitations of the LLM, we note that prompt engineering requires meticulous attention and testing to prevent repetitive dialogue patterns, especially when faced with content that is against the LLM's policy or the custom instructions. Similarly, using LLMs without proper constraints poses the risk of generating hallucinatory or offensive responses [6, 17]. Finally, the validity of our comparative results could be influenced by wizard expertise and as such, expanding beyond the two wizards we used would likely enhance the robustness of our findings.

5 CONCLUSION

In conclusion, our research confirms the potential of LLMs (GPT-4) in HRI, by showing that LLM-controlled robots are perceived as similarly socially intelligent as WoZ-controlled robots and are preferred by the participants for a brainstorming task. This insight opens new avenues for adopting and accepting autonomous robots with high social intelligence in real-world settings.

REFERENCES

[1] Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. 2012. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany, February 21-26, 2011, Revised Selected Papers*. Springer, 114–130. DOI: 10.1007/978-3-642-34584-5_9.

[2] Kimberly A. Barchard, Leiszle Lapping-Carr, R. Shane Westfall, Andrea Fink-Armold, Santosh Balajee Banisetty, and David Feil-Seifer. 2020. Measuring the perceived social intelligence of robots. *Journal of Human-Robot Interaction*, 9, 4, Article 24, (Sept. 2020), 29 pages. DOI: 10.1145/3415139.

[3] Erik Billing, Julia Rosén, and Maurice Lamb. 2023. Language models for human-robot interaction. In *ACM/IEEE International Conference on Human-Robot Interaction, March 13–16, 2023, Stockholm, Sweden*. ACM Digital Library, 905–906. DOI: 10.1145/3568294.3580040.

[4] Alwin de Rooij, Simone van den Broek, Michelle Bouw, and Jan de Wit. 2023. Co-designing with a social robot facilitator: effects of robot mood expression on human group dynamics. In *Proceedings of the 11th International Conference on Human-Agent Interaction* (HAI '23). Association for Computing Machinery, Gothenburg, Sweden, 22–29. DOI: 10.1145/3623809.3623820.

[5] Karen L. Dugosh and Paul B. Paulus. 2005. Cognitive and social comparison processes in brainstorming. *Journal of Experimental Social Psychology*, 41, 3, (May 1, 2005), 313–320. DOI: 10.1016/j.jesp.2004.05.009.

[6] Mohammadreza Farrokhnia, Seyyed Kazem Banihashem, Omid Noroozi, and A.E.J. Wals. 2023. A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, (Mar. 2023), 1–15. DOI: 10.1080/14703297.2023.2195846.

[7] Julia Geerts, Jan de Wit, and Alwin de Rooij. 2021. Brainstorming with a social robot facilitator: better than human facilitation due to reduced evaluation apprehension? *Frontiers in Robotics and AI*, 8, 657291. DOI: 10.3389/frobt.2021.657291.

[8] Bahar Irfan, Sanna-Mari Kuoppamäki, and Gabriel Skantze. 2023. Between reality and delusion: challenges of applying large language models to companion robots for open-domain dialogues with older adults. DOI: 10.21203/rs.3.rs-2884789/v1.

[9] Harold Jeffreys. 1998. *The Theory of Probability*. Oxford University Press.

[10] Prasanth Murali, Ian Steenstra, Hye Sun Yun, Ameneh Shamekhi, and Timothy Bickmore. 2023. Improving multiparty interactions with a robot using large language models. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI EA '23) Article 175. Association for Computing Machinery, 8 pages. DOI: 10.1145/3544549.3585602.

[11] Paul B. Paulus and Bernard A. Nijstad. 2019. *The Oxford Handbook of Group Creativity and Innovation*. Oxford Library of Psychology.

[12] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (HRI '15). Association for Computing Machinery, Portland, Oregon, USA, 141–148. DOI: 10.1145/2696454.2696497.

[13] Dominykas Strazdas, Jan Hintz, Anna-Maria Felßberg, and Ayoub Al-Hamadi. 2020. Robots and wizards: an investigation into natural human–robot interaction. *IEEE Access*, 8, 207635–207642. DOI: 10.1109/ACCESS.2020.3037724.

[14] Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. 2023. ChatGPT for robotics: design principles and model abilities. *Microsoft Autonomous Systems and Robotics Research*, 2, 20.

[15] Anita Vrins, Ethel Pruss, Jos Prinsen, Caterina Ceccato, and Maryam Alimardani. 2022. *Are You Paying Attention? The Effect of Embodied Interaction with an Adaptive Robot Tutor on User Engagement and Learning Performance*. (Jan. 2022), 135–145. DOI: 10.1007/978-3-031-24670-8_13.

[16] Eric-Jan Wagenmakers et al. 2018. Bayesian inference for psychology. Part II: example applications with JASP. *Psychonomic Bulletin & Review*, 25, 58–76. DOI: 10.3758/s13423-017-1323-7.

[17] Tom Williams, Cynthia Matuszek, Ross Mead, and Nick Depalma. 2023. Scarecrows in oz: the use of large language models in hri. *ACM Transactions on Human-Robot Interaction*. DOI: 10.1145/3606261.

[18] Ceng Zhang, Junxin Chen, Jiatong Li, Yanhong Peng, and Zebing Mao. 2023. Large language models for human–robot interaction: a review. *Biomimetic Intelligence and Robotics*, 3, 4, 100131. DOI: 10.1016/j.birob.2023.100131.