How Can Large Language Models Enable Better Socially Assistive Human-Robot Interaction: A Brief Survey

Zhonghao Shi, Ellen Landrum, Amy O'Connell, Mina Kian, Leticia Pinto-Alva, Kaleen Shrestha, Xiaoyuan Zhu, Maja J Matarić

University of Southern California, Los Angeles, United States
{zhonghas, elandrum, amy.dell, kian, pintoalv, kshresth, xzhu9839, mataric}@usc.edu

Abstract

Socially assistive robots (SARs) have shown great success in providing personalized cognitive-affective support for user populations with special needs such as older adults, children with autism spectrum disorder (ASD), and individuals with mental health challenges. The large body of work on SAR demonstrates its potential to provide at-home support that complements clinic-based interventions delivered by mental health professionals, making these interventions more effective and accessible. However, there are still several major technical challenges that hinder SAR-mediated interactions and interventions from reaching human-level social intelligence and efficacy. With the recent advances in large language models (LLMs), there is an increased potential for novel applications within the field of SAR that can significantly expand the current capabilities of SARs. However, incorporating LLMs introduces new risks and ethical concerns that have not yet been encountered, and must be carefully be addressed to safely deploy these more advanced systems. In this work, we aim to conduct a brief survey on the use of LLMs in SAR technologies, and discuss the potentials and risks of applying LLMs to the following three major technical challenges of SAR: 1) natural language dialog; 2) multimodal understanding; 3) LLMs as robot policies.

Introduction

At the intersection of assistive robots and socially interactive robots, *socially assistive robots* aim to provide assistance to support users' cognitive-affective needs through social interaction (Feil-Seifer and Matarić 2011). Past studies have demonstrated the benefits and potential of deploying SARs to support a diverse set of user populations, including individuals with mental health challenges (Scoglio et al. 2019), children with ASD (Cabibihan et al. 2013), older adults (Shishehgar, Kerr, and Blake 2018), and K-12 students (Randall 2019). With more affordable robot hardware (Koh, Ang, and Casey 2021; Pinto-Bernal et al. 2022; Koh et al. 2022), SARs have the potential to lower the socioeconomic barriers that limit access to personalized therapies, companionship, and education. However, prior work has also demonstrated that the existing SAR interactions have not yet been able to understand multimodal (language

and visual) social dynamics (Robinson et al. 2023; Li et al. 2021) and respond with human-like dialog (Skantze 2021) and actions (Akalin and Loutfi 2021).

With the recent advances in natural language processing (NLP) research, large language models (LLMs) such as GPT-4 have shown tremendous success in tasks both within the field of NLP (such as language modeling, question answering, and translation) (Achiam et al. 2023) and outside (such as programming (Xu et al. 2022), robot planning (Singh et al. 2023), and autonomous driving (Cui et al. 2024)). These capabilities of LLMs may open new possibilities for tackling the core technical challenges of SAR, and help us get closer to achieving more effective, human-level social assistance for users with differences.

In this work, we categorize the core technical challenges of SAR into three areas: 1) natural language dialog; 2) multimodal user understanding; 3) LLMs as robot policies. To survey the existing work on LLM-powered SARs in these three categories of technical challenges, we used Google Scholar to search the relevant papers in major human-robot interaction conferences, journals, and arXiv. In each of the following sections, we aim to: identify the potential of LLMs for SAR, survey existing work, and discuss future directions. This research was supported by the National Science Foundation Grant ITE-2236320 and IIS-1925083.

Natural Language Dialogue

Natural language dialogue is at the core of human-centered social interaction. Yet, prior to the recent breakthroughs in LLM technologies, SARs mainly relied on Wizard-of-Oz teleoperation (Erich, Hirokawa, and Suzuki 2017) or predefined rule-based dialogue management systems (Erich, Hirokawa, and Suzuki 2017; Youssef et al. 2022). SARs employing traditional non-LLM-based conversational systems are limited by their inability to accurately interpret human dialogue, limited vocabulary in dialogue generation, lack of understanding of context and ability to personalize, and lack of ability to effectively utilize online resources (Grassi, Recchiuto, and Sgorbissa 2022).

By applying state-of-art LLM models such as GPT-4 (Achiam et al. 2023), recent work on LLM-powered SAR has been mainly focused on enabling more accurate dialog understanding and more human-like and context-aware dialogue generation. LLM-powered SARs are able to pro-

duce varied dialogue while staying on topic (Billing, Rosén, and Lamb 2023). They can engage in more natural, flexible conversation with users from populations of interest, such as older adults and children with autism spectrum disorder (Bertacchini et al. 2023). Spitale, Axelsson, and Gunes (2023) designed an LLM-powered SAR as a motivational coach with both informative and emotional objectives, demonstrating that LLMs can be used to understand long-horizon context and enable long-term personalization. In instructional settings, LLM-powered SARs combine the vast knowledge base and interactive content-delivery of LLMs with the the capabilities and engaging nature of physically embodied agents (Wake et al. 2023).

Despite this progress, Irfan, Kuoppamäki, and Skantze (2023) showed that hallucinations, obsolete information, latency, and disengagement cues may still cause user frustration and confusion, which could be detrimental to socially assistive human-robot interaction. More exploration is needed to overcome these limitations and harness the power of LLMs in ways that better align with the goals of SAR interactions.

Multimodal User Understanding

To enable successful socially assistive human-robot interactions, a SAR needs to understand the user's cognitive-affective state (user engagement, affect, and intent) from multimodal perceptual data (language, visual, and audio) (Youssef et al. 2022). Existing work on multimodal social understanding has relied on training and fine-tuning machine learning (ML) models with data collected from previous SAR deployments (Robinson et al. 2023). However, the definition of cognitive-affective states may vary in different social contexts. Due to the independent and identically distributed (IID) assumption made by ML model training (Wang et al. 2022), existing ML models struggle to generalize effectively and quickly to test data that are distributed differently from the training data, particularly in the context of SAR (Shi et al. 2022).

Multimodal language models (MLMs) such as state-of-the-art vision-language models like CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021), and GPT-4V (Achiam et al. 2023), have shown promising zero-shot performance on a variety of human-centered visual tasks (Zhang et al. 2023a; Wu et al. 2023). Furthermore, these MLMs also demonstrate impressive few-shot capabilities of quickly adapting via prompting with natural language (Ge et al. 2023). This indicates that MLMs may also be capable of adapting to novel social context for more generalizable and accurate multimodal social understanding. Despite the recent progress in computer vision and robotics, using MLMs for SAR is still largely unexplored, but this direction of research shows great potential for significantly enabling better multimodal social understanding for socially assistive human-robot interaction.

LLMs as Robot Policies

In an ideal socially assistive human-robot interaction, a SAR should fluently learn and reason about the user's states and provide the best feedback or action as assistance. The space of user's states and robot feedback or actions can be large and continuous (Clabaugh and Matarić 2019) and existing approaches to encoding SAR policies, such as rule-based system and reinforcement learning, are not efficiently trained or sufficiently robust on large and continuous spaces with limited amounts of data (Akalin and Loutfi 2021). Past work has often circumvented this problem by constraining interactions to pre-defined tablet/computer games with small user state and action spaces (Clabaugh and Matarić 2019). LLMs may help to relax this constraint and enable more natural interaction by allowing continuous, more human-like formulation of space for user states and robot actions.

Recent studies in robotics and NLP have successfully employed LLMs as robot policies in the setting including autonomous driving (Cui et al. 2024), robot task planning (Singh et al. 2023; Ahn et al. 2022), social common sense (Sap et al. 2019) and social reasoning (Gandhi et al. 2023). In the field of SAR, existing work has mainly focused on applying LLMs to matching the affective state of robot feedback with the sentiment of the robot's dialog (Lee et al. 2023; Mishra et al. 2023; Lim et al. 2023). These studies have shown that LLM-powered robot policies for gesture matching enable more context-aware and natural robot gesture. Research using LLMs as robot policies has not yet explored 1) how to enable SARs to form robot policies for spontaneous tasks instead of only pre-defined ones, such as helping children with ASD navigate through an unpleasant social interaction they just encountered during their school day; 2) how to enable SAR policies to engage users with educational tasks while keeping them both challenged and encouraged; 3) how to reason socially about the user's intent and needs with partially observable information based on multimodal data; and 4) how to enable personalized SAR policies to quickly align with each user's unique needs, personality, and values, so the SAR can be more helpful and empathetic to each user.

Risks and Safety Considerations

Because SAR aims to support vulnerable populations including children and individuals with physical and/or mental health challenges (Matarić and Scassellati 2016), it is crucial to ensure absolute safety during socially assistive human-robot interactions. Despite the great success in LLMs, their lack of explainability and theoretical safety guarantees (Huang et al. 2023) may introduce significant risks and concerns by 1) amplifying unfairness and human bias (Acerbi and Stubbersfield 2023); 2) harming data security and privacy by unethical use of personal data (Liyanage et al. 2020); and 3) hallucination behaviors causing potential harms to users (Zhang et al. 2023b). For these reasons, the trustworthiness of LLM-powered SAR needs to be extensively evaluated before autonomous system can be deployed with vulnerable populations without human monitoring and oversight. As the first survey paper on this topic, this work aims to inform and stimulate research toward leveraging the tremendous potential of LLM-powered SARs and address the risks and safety concerns.

References

Acerbi, A.; and Stubbersfield, J. M. 2023. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44): e2313790120.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; Finn, C.; Fu, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Ho, D.; Hsu, J.; Ibarz, J.; Ichter, B.; Irpan, A.; Jang, E.; Ruano, R. J.; Jeffrey, K.; Jesmonth, S.; Joshi, N. J.; Julian, R.; Kalashnikov, D.; Kuang, Y.; Lee, K.-H.; Levine, S.; Lu, Y.; Luu, L.; Parada, C.; Pastor, P.; Quiambao, J.; Rao, K.; Rettinghouse, J.; Reyes, D.; Sermanet, P.; Sievers, N.; Tan, C.; Toshev, A.; Vanhoucke, V.; Xia, F.; Xiao, T.; Xu, P.; Xu, S.; Yan, M.; and Zeng, A. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. arXiv:2204.01691.

Akalin, N.; and Loutfi, A. 2021. Reinforcement learning approaches in social robotics. *Sensors*, 21(4): 1292.

Bertacchini, F.; Demarco, F.; Scuro, C.; Pantano, P.; and Bilotta, E. 2023. A social robot connected with chatGPT to improve cognitive functioning in ASD subjects. *Frontiers in Psychology*, 14.

Billing, E.; Rosén, J.; and Lamb, M. 2023. Language models for human-robot interaction. In *ACM/IEEE International Conference on Human-Robot Interaction, March 13–16, 2023, Stockholm, Sweden*, 905–906. ACM Digital Library.

Cabibihan, J.-J.; Javed, H.; Ang, M.; and Aljunied, S. M. 2013. Why robots? A survey on the roles and benefits of social robots in the therapy of children with autism. *International journal of social robotics*, 5: 593–618.

Clabaugh, C.; and Matarić, M. 2019. Escaping oz: Autonomy in socially assistive robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 2: 33–61.

Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Zhou, Y.; Liang, K.; Chen, J.; Lu, J.; Yang, Z.; Liao, K.-D.; et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 958–979.

Erich, F.; Hirokawa, M.; and Suzuki, K. 2017. A systematic literature review of experiments in socially assistive robotics using humanoid robots. *arXiv preprint arXiv:1711.05379*.

Feil-Seifer, D.; and Matarić, M. J. 2011. Socially assistive robotics. *IEEE Robotics & Automation Magazine*, 18(1): 24–31.

Gandhi, K.; Fränken, J.-P.; Gerstenberg, T.; and Goodman, N. D. 2023. Understanding social reasoning in language models with language models. *arXiv preprint arXiv:2306.15448*.

Ge, W.; Chen, S.; Chen, G.; Chen, J.; Chen, Z.; Yan, S.; Zhu, C.; Lin, Z.; Xie, W.; Wang, X.; et al. 2023. MLLM-Bench, Evaluating Multi-modal LLMs using GPT-4V. *arXiv preprint arXiv:2311.13951*.

Grassi, L.; Recchiuto, C. T.; and Sgorbissa, A. 2022. Knowledge-grounded dialogue flow management for social robots and conversational agents. *International Journal of Social Robotics*, 14(5): 1273–1293.

Huang, X.; Ruan, W.; Huang, W.; Jin, G.; Dong, Y.; Wu, C.; Bensalem, S.; Mu, R.; Qi, Y.; Zhao, X.; et al. 2023. A Survey of Safety and Trustworthiness of Large Language Models through the Lens of Verification and Validation. *arXiv preprint arXiv:2305.11391*.

Irfan, B.; Kuoppamäki, S.-M.; and Skantze, G. 2023. Between Reality and Delusion: Challenges of Applying Large Language Models to Companion Robots for Open-Domain Dialogues with Older Adults.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Koh, W. Q.; Ang, F. X. H.; and Casey, D. 2021. Impacts of low-cost robotic pets for older adults and people with dementia: scoping review. *JMIR rehabilitation and assistive technologies*, 8(1): e25340.

Koh, W. Q.; Whelan, S.; Heins, P.; Casey, D.; Toomey, E.; and Dröes, R.-M. 2022. The usability and impact of a low-cost pet robot for older adults and people with dementia: qualitative content analysis of user experiences and perceptions on consumer websites. *JMIR aging*, 5(1): e29224.

Lee, Y. K.; Jung, Y.; Kang, G.; and Hahn, S. 2023. Developing Social Robots with Empathetic Non-Verbal Cues Using Large Language Models. *arXiv preprint arXiv:2308.16529*.

Li, Z.; Mu, Y.; Sun, Z.; Song, S.; Su, J.; and Zhang, J. 2021. Intention understanding in human–robot interaction based on visual-NLP semantics. *Frontiers in Neurorobotics*, 14: 610139.

Lim, J.; Sa, I.; MacDonald, B.; and Ahn, H. S. 2023. A Sign Language Recognition System with Pepper, Lightweight-Transformer, and LLM. *arXiv preprint arXiv:2309.16898*.

Liyanage, H.; Kuziemsky, C.; Terry, A.; Schreiber, R.; Jonnagaddala, J.; de Lusignan, S.; McGovern, A.; Hinton, W.; Corra, A.; Munro, N.; et al. 2020. Ethical use of electronic health record data and artificial intelligence: recommendations of the primary care informatics working group of the international medical informatics association. *Yearbook of medical informatics*, 29(01): 051–057.

Matarić, M. J.; and Scassellati, B. 2016. Socially assistive robotics. *Springer handbook of robotics*, 1973–1994.

Mishra, C.; Verdonschot, R.; Hagoort, P.; and Skantze, G. 2023. Real-time emotion generation in human-robot dialogue using large language models. *Frontiers in Robotics and AI*, 10.

Pinto-Bernal, M. J.; Cespedes, N.; Castro, P.; Munera, M.; and Cifuentes, C. A. 2022. Physical human-robot interaction influence in ASD therapy through an affordable soft social robot. *Journal of Intelligent & Robotic Systems*, 105(3): 67.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.;

et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Randall, N. 2019. A survey of robot-assisted language learning (RALL). *ACM Transactions on Human-Robot Interaction (THRI)*, 9(1): 1–36.

Robinson, N.; Tidd, B.; Campbell, D.; Kulić, D.; and Corke, P. 2023. Robotic vision for human-robot interaction and collaboration: A survey and systematic review. *ACM Transactions on Human-Robot Interaction*, 12(1): 1–66.

Sap, M.; Rashkin, H.; Chen, D.; Le Bras, R.; and Choi, Y. 2019. Social IQa: Commonsense Reasoning about Social Interactions. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4463–4473. Hong Kong, China: Association for Computational Linguistics.

Scoglio, A. A.; Reilly, E. D.; Gorman, J. A.; and Drebing, C. E. 2019. Use of social robots in mental health and well-being research: systematic review. *Journal of medical Internet research*, 21(7): e13322.

Shi, Z.; Groechel, T. R.; Jain, S.; Chima, K.; Rudovic, O.; and Matarić, M. J. 2022. Toward personalized affect-aware socially assistive robot tutors for long-term interventions with children with autism. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(4): 1–28.

Shishehgar, M.; Kerr, D.; and Blake, J. 2018. A systematic review of research into how robotic technology can help older people. *Smart Health*, 7: 1–18.

Singh, I.; Blukis, V.; Mousavian, A.; Goyal, A.; Xu, D.; Tremblay, J.; Fox, D.; Thomason, J.; and Garg, A. 2023. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 11523–11530. IEEE.

Skantze, G. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*, 67: 101178.

Spitale, M.; Axelsson, M.; and Gunes, H. 2023. VITA: A Multi-modal LLM-based System for Longitudinal, Autonomous, and Adaptive Robotic Mental Well-being Coaching. *arXiv preprint arXiv:2312.09740*.

Wake, N.; Kanehira, A.; Sasabuchi, K.; Takamatsu, J.; and Ikeuchi, K. 2023. GPT Models Meet Robotic Applications: Co-Speech Gesturing Chat System. *arXiv preprint arXiv:2306.01741*.

Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; and Yu, P. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*.

Wu, W.; Yao, H.; Zhang, M.; Song, Y.; Ouyang, W.; and Wang, J. 2023. GPT4Vis: What Can GPT-4 Do for Zero-shot Visual Recognition? *arXiv preprint arXiv:2311.15732*.

Xu, F. F.; Alon, U.; Neubig, G.; and Hellendoorn, V. J. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, 1–10.

Youssef, K.; Said, S.; Alkork, S.; and Beyrouthy, T. 2022. A survey on recent advances in social robotics. *Robotics*, 11(4): 75.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2023a. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. 2023b. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv preprint arXiv:2309.01219*.