A Multi-party Conversational Social Robot Using LLMs

Angus Addlesee
Heriot-Watt University
Edinburgh, UK
a.addlesee@hw.ac.uk

Neeraj Cherakara
Heriot-Watt University
Edinburgh, UK
nc2025@hw.ac.uk

Nivan Nelson
Heriot-Watt University
Edinburgh, UK
nn2023@hw.ac.uk

Daniel Hernández García
Heriot-Watt University
Edinburgh, UK
d.hernandez_garcia@hw.ac.uk

Nancie Gunson
Heriot-Watt University
Edinburgh, UK
n.gunson@hw.ac.uk

Weronika Sieińska
Heriot-Watt University
Edinburgh, UK
w.sieinska@hw.ac.uk

Marta Romeo
Heriot-Watt University
Edinburgh, UK
m.romeo@hw.ac.uk

Christian Dondrup
Heriot-Watt University
Edinburgh, UK
c.dondrup@hw.ac.uk

Oliver Lemon
Heriot-Watt University
Edinburgh, UK
o.lemon@hw.ac.uk

ABSTRACT

In this paper, we describe our setting and the architecture of our LLM-based dialogue system embodied in a social robot and able to have multi-party conversations. Each component is detailed, and a video of the full system is available with the appropriate components highlighted in real-time. Our system decides when it should take its turn, generates human-like clarification requests when the patient pauses mid-utterance, answers in-domain questions (grounding to the in-prompt knowledge), and responds appropriately to out-of-domain requests (like generating jokes or quizzes).

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; *Accessibility technologies*; **Natural language interfaces**; *Sound-based input / output*.

KEYWORDS

social robots, human-robot interaction, conversational AI, accessibility, multi-party dialogue, large language models

ACM Reference Format:
Angus Addlesee, Neeraj Cherakara, Nivan Nelson, Daniel Hernández García, Nancie Gunson, Weronika Sieińska, Marta Romeo, Christian Dondrup, and Oliver Lemon. 2024. A Multi-party Conversational Social Robot Using LLMs. 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, 3 pages. https://doi.org/10.1145/3610978.3641112

1 INTRODUCTION

Both commercial and research spoken dialogue systems (SDSs), conversational agents, and social robots have been designed with a focus on dyadic interactions. That is, a two-party conversation between one individual user and a single system/robot. These are only guaranteed in specific settings, like people interacting with Siri on their own phone, or with Amazon Alexa in single-occupant homes. In a family home, Alexa's lack of multi-party capabilities are apparent [15], but this becomes a critical limitation when deploying social robots in public spaces. Families visit museums and libraries, groups of friends roam shopping malls and bars, and couples travel through airports and support each other at hospital appointments. Social robots are being tested in all of these settings [3, 8, 9, 13, 17], in which multi-party conversations (MPCs), involving people talking to both the robot and each other, commonly occur.

Ex	User	Utterance	Note of Interest
(A)	U1 U1	I think it is London Yeah... London	If turn 2 was U2, it would be agreement, so speaker recognition changes meaning.
(B)	U1	My dad needs the toilet	Providing other user's goal.
(C)	U1 U2	When is my scan? It's at 10am	U2 answers U1's question, but addressee was ambiguous without gaze info.
(D)	U1 ARI	We are hungry The café is to the left, but you should fast after 10am	Shared goal indicated by 'we', and robot can point to the 'left'. Fasting is in red as it is a world-knowledge hallucination.
(E)	U1 ARI U1 ARI	Name a song by... By who? Queen Bohemian Rhapsody	This is an OOD question that could not be answered without the LLM-based SDS. The partial utterance is handled naturally which improves accessibility.

Table 1: Example multi-party conversations.

Tasks that are typically trivial in the dyadic setting become considerably more complex when conversing with multiple users [10, 16]: (1) The speaker is no longer simply the other person, so the meaning of the dialogue depends on recognising who said each utterance (see (A) in Table 1); (2) similarly, addressee recognition becomes more complicated (see Sec 3) as people can address each other, the robot, and groups; and (3) response generation depends on who said what to whom, relying on the semantic content and surrounding multi-party context. To make things even more difficult, MPCs provide additional unique challenges that are underexplored. Dyadic SDSs must identify and answer the user's goals to be practically useful. In MPCs, users can provide another person's goal (see (B) in Table 1), answer each other's goals (see (C) in Table 1), and even share goals (see (D) in Table 1, [7]), which we find in our setting [2].

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

ACM ISBN 979-8-4007-0323-2/24/03
https://doi.org/10.1145/3610978.3641112

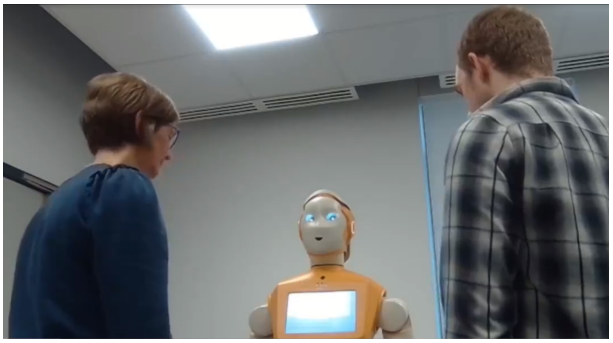

Figure 1: A screenshot of the video described in this paper presenting a multi-party conversation with the robot. The video has been shot in the lab of our university and it is a representation of the system currently under testing in a memory clinic.

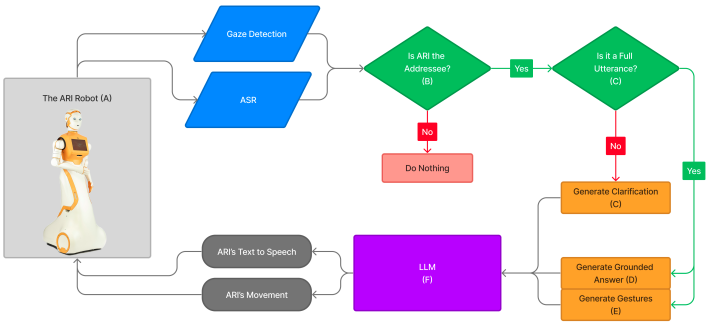

Figure 2: The architecture of our multi-party dialogue system deployed on the ARI robot.

Large language models (LLMs) have revolutionised our field. They are excellent at language understanding, and this includes MPCs [18] as their pre-training includes scripts and meeting transcripts containing multiple people. They hold a wealth of general knowledge, enabling abilities like question answering (QA) and playing quizzes.

2 SETTING

The work described herein is being carried out in the context of robots deployed in gerontological healthcare clinics, as part of the EU H2020-ICT funded SPRING project, to provide assistance to patients and visitors. Our goal is to provide a system that is both practically useful, but also entertaining, to provide participants with some light distraction from their otherwise stressful day.

In this paper, we describe current progress towards developing an LLM-based multi-party conversational AI system integrated in a Social Conversational Robot (an ARI robot [5]) that will act as a receptionist in a hospital waiting room. Our system has been iteratively improved through regular user tests, including patients visiting a hospital memory clinic. A multi-party data collection has been designed and conducted [1, 2], and this data has been used to motivate and evaluate the system we present here.

An initial system [11] was developed before the recent LLM advance, relying on a 'traditional' modular architecture based upon Alana V2 [6]. The lack of multi-party capabilities of that system proved problematic. It interrupted user's conversations, as it responded to every turn, not allowing them to talk to each other at any point. Our new multi-party LLM-based SDS, embodied by the ARI social robot, not only is multi-party, but it also improves QA accuracy, improves accessibility, and enables added functionality. For example, where previously, we had to specifically design the system to tell jokes and run entertaining quizzes, LLMs can now handle this inherently due to their world knowledge.

3 DIALOGUE SYSTEM

In this section, we detail each system component in Figure 2. All prompts used in the system can be found on GitHub[1].

Detecting the User's Addressee: Unlike dyadic systems, that will reply to every user's turn, people talk to both the robot and each other in MPCs. For example, if the robot replied to U1 in Example (C), Table 1, then it would have interrupted U2. However, the addressee of U1's turn is ambiguous given the text alone. An addressee detection prompt was developed to ask the LLM whether the user utterance "is currently addressing the other person or the robot" based on the content of the sentence and a signal about the current speaker's gaze, which can be provided by a gaze detection module. We first classify the user utterances as whether the robot is the addressee or not. When the robot is not being addressed by the user utterance, we simply ignore it (pass the turn) and continue to listen without interrupting the conversation.

[1]https://github.com/AddleseeHQ/mp-llm-demo-prompts

Generating Clarification Requests: In a hospital's memory clinic, voice accessibility is critical. Pauses are easily mistaken as end of turn by the ASR, resulting in the user being interrupted with nonsense or a generic response like "I didn't understand that". The user is forced to repeat their entire turn again, a frustrating and unnatural interaction [12, 14].

Accessibility settings, in Siri for example, allow users to modify how long the ASR waits until it decides that a sentence is complete. This is a wonderful temporary solution for people with more progressed cognitive impairment, but it is not naturally interactive, as the user would then have to wait for long durations between *every* turn. Producing incremental clarification requests (iCRs) is, therefore, important for building naturally interactive SDSs [4]. In order to handle our user's incomplete sentences, we first ask the LLM whether the turn was a complete sentence. If it is not, we generate an iCR by giving it examples from a human iCR corpus[2]. This can be seen in the architecture in Figure 2, denoted by (C).

Generating Responses: One huge benefit of using LLMs is their inherent ability to perform general chit-chat, tell jokes, and access a wealth of general knowledge. In the original system, we could only respond suitably to utterances that the system was pre-designed to handle, and we would attempt to respond to unexpected utterances with suggestions (e.g. "I'm not sure, but I can help you with directions."). Unexpected utterances can now be handled directly by the LLM. However, SDS based on LLMs face new challenges. For example, LLMs could disclose personal or harmful information. For this reason, We provide the hospital information to our LLM in a prompt with some additional guardrails, like "you are not qualified to give any medical advice or make medical diagnoses" and "you do not have access to individual patient records or schedules". In addition, the LLM's static world knowledge can cause harmful hallucinations due to conflicts with the information given in the prompt. The text in red in Example (D) in Table 1 highlights this issue. Our prompt does not state that patients must fast after 10am, and this response would result in a patient going hungry. We found that a prompt providing the passage as a quote by a fictitious non-celebrity name, "Jodie W. Jenkins", grounded responses to the in-prompt knowledge. We ask the LLM to answer according to 'Jodie', and our SDS utilises this prompt in component (D) in Figure 2.

4 CONCLUSIONS AND FUTURE WORK

Our SDS is embodied by the ARI social robot. Using data collected with real memory clinic patients in this complex setting, our system is able to decide when to take its turn, generate natural clarification requests to improve accessibility, answer in-domain questions grounded to our domain-specific knowledge, and respond appropriately to out-of-domain requests like generating jokes, quizzes, and general chit-chat. We are currently running further data collection in the hospital with the LLM-based SDS. Using this data, we will further refine our system and curate corpora that will be released to allow other researchers to work on this complex, yet vital task. In addition, we are working on producing helpful gestures exploiting the motion capabilities of the robot. We plan to generate and integrate context-dependent gestures, as illustrated through component (E) in Figure 2, using the Vicuna-13b-v1.5 LLM (component (F)) in parallel with the grounded answer generation.

[2]https://github.com/AddleseeHQ/interruption-recovery

ACKNOWLEDGMENTS

This research was funded by the EU H2020 program under grant agreement no. 871245 (https://spring-h2020.eu/).

REFERENCES

[1] Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernández Garcia, Christian Dondrup, and Oliver Lemon. 2023. Data Collection for Multi-party Task-based Dialogue in Social Robotics. In *Proceedings of the 13th International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

[2] Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernández García, Christian Dondrup, and Oliver Lemon. 2023. Multi-party Goal Tracking with LLMs: Comparing Pre-training, Fine-tuning, and Prompt Engineering. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. https://sigdialinlg2023.github.io/paper_sigdial109.html

[3] 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.

[4] Javier Chiyah-Garcia, Alessandro Suglia, Arash Eshghi, and Helen Hastie. 2023. 'What are you referring to?' Evaluating the Ability of Multi-Modal Dialogue Models to Process Clarificational Exchanges. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czechia, 175–182. https://aclanthology.org/2023.sigdial-1.16

[5] Sara Cooper, Alessandro Di Fava, Carlos Vivas, Luca Marchionni, and Francesco Ferro. 2020. ARI: The Social Assistive Robot and Companion. In *29th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2020*. 745–751.

[6] Amanda Cercas Curry, Ioannis Papaioannou, Alessandro Suglia, Shubham Agarwal, Igor Shalyminov, Xinnuo Xu, Ondrej Dušek, Arash Eshghi, Ioannis Konstas, Verena Rieser, et al. 2018. Alana v2: Entertaining and informative open-domain social dialogue using ontologies and entity linking. *Alexa Prize* (2018).

[7] Arash Eshghi and Patrick GT Healey. 2016. Collective contexts in conversation: Grounding by proxy. *Cognitive science* 40, 2 (2016), 299–324.

[8] Mary Ellen Foster, Bart Craenen, Amol Deshmukh, Oliver Lemon, Emanuele Bastianelli, Christian Dondrup, Ioannis Papaioannou, Andrea Vanzo, Jean-Marc Odobez, Olivier Canévet, et al. 2019. MuMMER: Socially intelligent human-robot interaction in public spaces. *arXiv preprint arXiv:1909.06749* (2019).

[9] Press Furhat Robotics. 2015. FRAnny, Frankfurt Airport's new multilingual robot concierge can help you in over 35 languages. *Furhat Robotics Press Release* (May 2015). https://furhatrobotics.com/press-releases/franny-frankfurt-airports-new-multilingual-robot-concierge-can-help-you-in-over-35-languages/

[10] Jia-Chen Gu, Chongyang Tao, and Zhen-Hua Ling. 2022. WHO Says WHAT to WHOM: A Survey of Multi-Party Conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*.

[11] Nancie Gunson, Daniel Hernández García, Weronika Sieińska, Christian Dondrup, and Oliver Lemon. 2022. Developing a Social Conversational Robot for the Hospital waiting room. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1352–1357.

[12] Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How do users respond to voice input errors? Lexical and phonetic query reformulation in voice search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 143–152.

[13] Simon Keizer, Mary Ellen Foster, Zhuoran Wang, and Oliver Lemon. 2014. Machine learning for social multiparty human–robot interaction. *ACM transactions on interactive intelligent systems (TIIS)* 4, 3 (2014), 1–32.

[14] Mikio Nakano, Yuka Nagano, Kotaro Funakoshi, Toshihiko Ito, Kenji Araki, Yuji Hasegawa, and Hiroshi Tsujino. 2007. Analysis of user reactions to turn-taking failures in spoken dialogue systems. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*. 120–123.

[15] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[16] David Traum. 2004. Issues in multiparty dialogues. In *Advances in Agent Communication: International Workshop on Agent Communication Languages, ACL 2003, Melbourne, Australia, July 14, 2003. Revised and Invited Papers*. Springer.

[17] Evgenios Vlachos, Anne Faber Hansen, and Jakob Povl Holck. 2020. A robot in the library. In *International conference on human-computer interaction*. Springer.

[18] Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. DialogLM: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36.