# Multi-party Multimodal Conversations Between Patients, Their Companions, and a Social Robot in a Hospital Memory Clinic

**Angus Addlesee, Neeraj Cherakara, Nivan Nelson, Daniel Hernández García**
**Nancie Gunson, Weronika Sieińska, Christian Dondrup, Oliver Lemon**
The Interaction Lab
Heriot-Watt University
Edinburgh, UK

## Abstract

We have deployed an LLM-based spoken dialogue system in a real hospital. The ARI social robot embodies our system, which patients and their companions can have multi-party conversations with together. In order to enable this multi-party ability, multimodality is critical. Our system, therefore, receives speech and video as input, and generates both speech and gestures (arm, head, and eye movements). In this paper, we describe our complex setting and the architecture of our dialogue system. 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). This latter feature is particularly remarkable as real patients often utter unexpected sentences that could not be handled previously.

## 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. When Alexa is in a family home, their lack of multi-party capabilities are apparent (Porcheron et al., 2018), 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 deployed and tested in all of these settings (Al Moubayed et al., 2012; Keizer et al., 2014; Furhat Robotics, 2015; Foster et al., 2019; Vlachos et al., 2020; Gunson et al., 2022), in which multi-party conversations (MPCs), involving people talking to both the robot and each other, do commonly occur (see Figure 1).

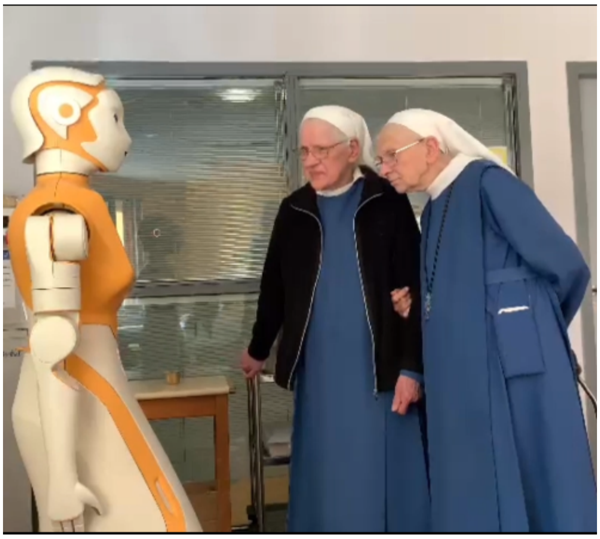

Figure 1: Hospital memory clinic visitors using our SDS on the ARI social robot (Cooper et al., 2020).

Tasks that are typically trivial in the dyadic setting become considerably more complex when conversing with multiple users (Traum, 2004; Gu et al., 2022b): (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) addressee recognition is similarly more complicated (see Sec 3.2) as people 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

*Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*
*System Demonstrations*, pages 62–70
March 17-22, 2024 ©2024 Association for Computational Linguistics

| Example | User | Utterance | Note of Interest |
|---|---|---|---|
| (A) | U1<br>U1 | I think it is London<br>Yeah... London | If turn 2 was U2, it would be agreement, so speaker recognition changes meaning. |
| (B) | U1 | My husband needs the bathroom | Providing other user's goal. |
| (C) | U1<br>U2 | What time is my appointment?<br>It's at 10am | U2 answers U1's question, but addressee was ambiguous without gaze info. |
| (D) | U1<br>ARI | We are hungry<br>The café is through the door on your left, but you should fast before your visit. | 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<br>ARI<br>U1<br>ARI | Name a song by...<br>By who?<br>Queen<br>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: Utterances and interactions that illustrate behaviours of interest to this paper (referred to where appropriate). Examples B & C from MPCs with hospital memory clinic patients, their companions, and our SDS on the ARI robot. Example A: (Schauer et al., 2023). Examples D & E: (Addlesee, 2024).

each other's goals (see (C) in Table 1), and even share goals (see (D) in Table 1, (Eshghi and Healey, 2016)). We therefore established multi-party goal-tracking in previous work (Addlesee et al., 2023d).

Both dyadic and multi-party human conversations are subtly guided and supported by visual cues (Goodwin, 1981; Bavelas and Gerwing, 2011; Addlesee et al., 2019). Screwing-up of the face, brow furrows, looking up, nodding, smiling, eye-contact, etc... though crucial, are lost completely by current commercial SDSs. Due to the added complexity of MPCs, visual cues are even more crucial (Moujahid et al., 2022). For example, It is ambiguous who U1 is addressing in Example (C) in Table 1 because gaze behaviour is essential (Auer, 2018), yet missing.

In this paper, we present our multi-party multimodal SDS embodied by the ARI social robot (Cooper et al., 2020) that is currently deployed in a hospital, and interacts with memory clinic patients and their companions. It can give directions, provide light entertainment (like quizzes and jokes), and inform people about bus times, the cafe menu, and more. Large language models (LLMs) have revolutionised our field, they are excellent at language understanding, and this includes MPCs (Hu et al., 2019; Gu et al., 2021, 2022a; Zhong et al., 2022) as their pre-training includes scripts and meeting transcripts containing multiple people. They also hold a wealth of general knowledge, enabling abilities like question answering (QA), joke telling, and playing quizzes. Our SDS is therefore LLM-based to provide a state-of-the-art experience for hospital patients. We first describe our setting, and then detail each module of our system's architecture in Figure 2. A demo video of this system is available on YouTube[1].

## 2 The Hospital Setting

Dementia diagnosis is a stressful process. Patients typically spend entire days at the hospital with a friend or family member for support. The hours are filled with multiple appointments, but a large portion of the day is also spent waiting anxiously for test results or the next appointment. 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. The research staff at the hospital are our collaborators on the SPRING project, and they run the experiments with volunteer patients, their companions, and the ARI robot (see Figure 1).

The EU's H2020 SPRING project aims to explore "how to create robots able to move, see, hear and communicate with several actors, in complex and unstructured populated spaces"[2]. We are one of eight project partners, and our focus is the SDS. Other partners work on collision prevention during navigation, route planning, ego-noise suppression, gaze tracking, running live experiments with patients in the hospital, and more.

## 3 Dialogue System

Our system presented in this paper has been iteratively improved through regular user tests and interviews with patients visiting the hospital memory clinic. The initial system (Gunson et al., 2022) was developed before the recent LLM advance, relying on a 'traditional' modular architecture based upon

[1] https://www.youtube.com/watch?v=xMCpcsLhN_I

[2] https://spring-h2020.eu/

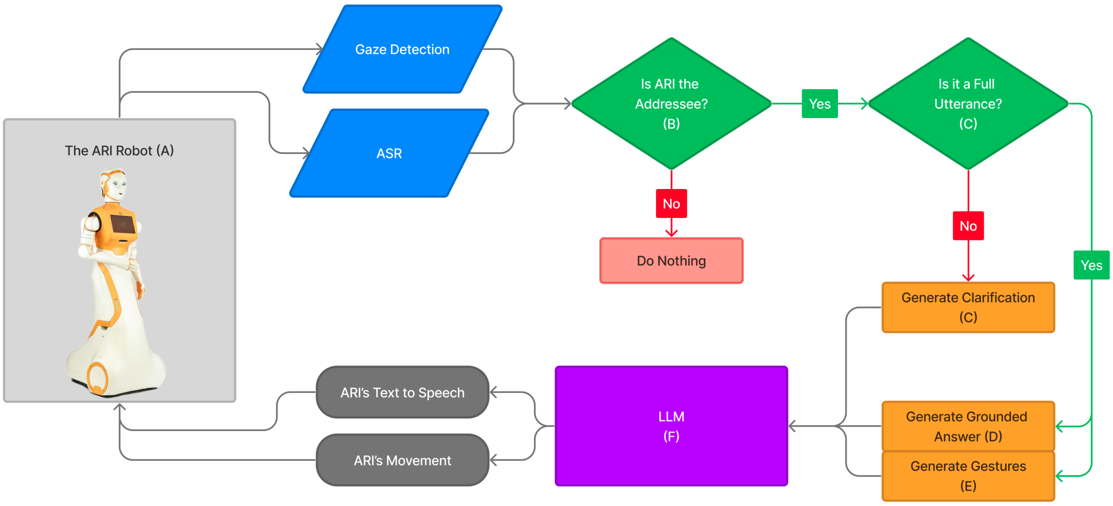

Figure 2: The architecture of our multi-party multimodal dialogue system deployed on the ARI robot.

Alana V2 (Papaioannou et al., 2017; Curry et al., 2018). As patients were usually accompanied by a companion, the lack of multi-party capabilities proved problematic. It interrupted users as it responded to every turn, not allowing them to talk to each other at any point. We therefore designed and ran a multi-party data collection in a wizard-of-oz setup (Addlesee et al., 2023c,d), and have used this data to motivate and evaluate the system we present here. Not only is this new system multi-party and multimodal, it improves QA accuracy, improves accessibility to people with dementia (Addlesee, 2024), and enables added functionality. Where previously, we had to specifically design the system to tell jokes and run entertaining quizzes (Addlesee et al., 2023a; Schauer et al., 2023), LLMs can now handle this inherently due to their world knowledge. Most importantly, both users and the hospital staff have reported that the user experience has improved drastically. In this section, we detail each system component illustrated in Figure 2.

### 3.1 Robot Platform

Our system is deployed on the ARI humanoid robot, designed for use as a socially assistive companion (Cooper et al., 2020). ARI is 1.65m tall, has a mobile base, a touch-screen on the torso, movable arms to gesture, and a head with LCD eyes that enable gaze behaviour. A photo of ARI can be seen in Figure 1 and component (A) in Figure 2. It is equipped with a ReSpeaker Mic v2.0 array[3], an RGB camera (in the head), and a 180º fish-eye camera (in the chest) allowing us to capture and record the audio and video of the whole interaction from the robot's perspective. The robot verbalises given responses using Acapela Text-To-Speech[4].

### 3.2 Detecting the User's Addressee

Dyadic SDSs reply to every user turn. As discussed in Section 1, people talk to both the robot and each other in MPCs. If the robot replied to U1 in Example (C), Table 1, then it would have interrupted U2. The addressee of U1's turn is ambiguous given the text alone. Alternatively, if the user said "Do you want to sit down?", it would be clear that ARI is not being addressed from just the text. In order to measure how effective gaze information is to determine the addressee in our specific setting, we annotated real MPCs collected in the hospital. We have video recordings of the interactions with the robot's cameras and an external camera. Using both the video and audio, the gold addressee of each turn was annotated along with whether the user was looking at ARI or not.

Using the Vicuna-13b-v1.5 LLM (Chiang et al., 2023), we created two addressee detectors. In one case, we prompted it with the dialogue history and current user's turn. In the second case, we added whether the user is looking at ARI or not. Both prompts asked the LLM whether the user "is currently addressing the other person or the robot"[5].

Addressee detection accuracy increased from 53.35% to 85.40% when given the gaze information. Reducing interruption of the user is a huge

[3] https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0

[4] https://www.acapela-group.com/

[5] All the exact prompts can be found on GitHub https://github.com/AddleseeHQ/mp-llm-demo-prompts.

improvement, but we do not want the robot to start ignoring people entirely. That is, we do not want the patient to address the robot and get no response. It is therefore critical to maximise recall, which increased from 31.33% to 91.00% when provided gaze information. A gaze detection model (Tonini et al., 2023) is used to get information on when a speaker is looking at ARI, and this is fed into component (B) in Figure 2.

### 3.3 Generating Clarification Requests

In a hospital's memory clinic, voice accessibility is critical (Addlesee, 2023), and people with dementia pause more frequently and for longer durations mid-sentence due to word-finding problems (Boschi et al., 2017; Slegers et al., 2018). These pauses are mistaken as end of turn by the ASR, resulting in the user being interrupted with nonsense or a generic response like "I'm sorry, I didn't understand that". The user is forced to repeat their entire turn again, a frustrating and unnatural interaction (Nakano et al., 2007; Jiang et al., 2013; Panfili et al., 2021).

Accessibility settings, in Siri for example (Apple, 2022), 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 (Chiyah-Garcia et al., 2023).

#### 3.3.1 CR Corpus and Taxonomy

Corpora of interrupted sentences paired with their meaning representations were recently released (Addlesee and Damonte, 2023a,b), finding that interrupted sentence recovery pipelines reliant on CRs were best at recovering the intended meaning of the question. They did not focus on generating natural, human-like iCRs in response to partial sentences. Using a subset of their SLUICE corpus (Addlesee and Damonte, 2023a), we elicited 12 CRs from annotators for 250 interrupted questions. This new corpus SLUICE-CR, therefore, contains a total of 3,000 human CRs (Addlesee, 2024).

All CRs within SLUICE-CR are intended to elicit how the interlocutor would have gone on to complete their turn. Example (E) in Table 1 illustrates this. Each CR in the corpus is classified into one of four distinct categories. First, there are sentential CRs (SentCRs), and these stand on their own as full sentences (e.g. "Who wrote what?"). We can see in Table 2 that humans rarely generated these, but LLMs that were not exposed to SLUICE-CR (the basic prompt) relied predominantly on SentCRs. All other CRs in the corpus are iCRs, fragments that are constructed as a continuation or completion of the truncated turn. iCRs are classified further. Reprise CRs (RCRs) simply retrace some of the words from the end of the truncated turn to localise the point of interruption (Howes et al., 2012), for example, responding "zipcode of?" in response to "What is the zipcode of...". Sluice CRs (SCRs) are similar to RCRs, but they end in a wh-word (who, what, where, etc...). For example, "zipcode of who?" or Example (E) in Table 1.

Table 2: Clarification request generation results. SMA: Sluice Match Accuracy. SentCR: Sentential CR. RCR: Reprise CR. SCR: Sluice CR. Prompt styles = Basic, Annotation, and Reasoning.

| Model | Prompt | SMA | SentCR | RCR | SCR |
|---|---|---|---|---|---|
| Human | - | - | 3.8 | 39.6 | 35.2 |
| GPT-4 | B | 11.7 | 91.2 | 0.0 | 0.0 |
| | A | **98.4** | 6.8 | 1.2 | 79.6 |
| | R | 97.6 | 0.8 | 1.2 | 86.0 |
| Llama-2 13b-chat | B | 3.3 | 91.6 | 0.4 | 0.0 |
| | A | 0.0 | 100 | 0.0 | 0.0 |
| | R | 2.0 | 99.2 | 0.0 | 0.0 |
| Llama-2 70b-chat | B | 2.6 | 99.6 | 0.0 | 0.0 |
| | A | 91.6 | 69.2 | 7.6 | 8.4 |
| | R | 86.0 | 51.6 | 20.0 | 12.0 |
| Vicuna 13b-v1.5 | B | 11.7 | 98.4 | 0.0 | 0.0 |
| | A | 83.9 | 73.2 | 0.0 | 20.4 |
| | R | 87.0 | 66.4 | 2.4 | 20.0 |

#### 3.3.2 CR Results

With that taxonomy in mind, we evaluated LLMs using SLUICE-CR (Addlesee, 2024). The results relevant to the hospital deployment can be found in Table 2. The 'basic' prompt simply passed the truncated turn to each LLM with no context. The 'annotation' prompt contained the task instructions given to the human annotators, which contains CR examples, and the 'reasoning' prompt added a reason for each example (Fu et al., 2022).

Of the models that learned to generate iCRs, GPT-4 and Vicuna-13b-v1.5 both relied more on SCRs. Llama-70b-chat generated more RCRs, opting to commonly forego the sluice entirely. Generating human-like iCRs is practically useless if they are not semantically appropriate. 85.5% of the human CRs contained a sluice, so we devised a new metric called the sluice match accuracy (SMA): measuring the percentage of model generated CRs with a wh-word that is an exact match to at least one of the wh-words in the 12 human CRs for each partial question. SMA thereby preserves semantic type ambiguity captured by the human-annotators.

From these metrics alone, it is clear that GPT-4 is outstanding if data privacy is not a concern. In sensitive settings without hardware limitations, Llama-2-70b-chat is best. Given our sensitive setting with hardware limitations, we use Vicuna-13b-v1.5 as our system's core LLM. 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 use the 'reasoning' prompt to generate an iCR to create a more accessible and naturally interactive conversational system. This can be seen in the architecture in Figure 2, denoted by (C).

### 3.4 Generating Responses

Unlike older dialogue systems, we interface with our core LLM using prompts. As mentioned in Section 3.3, we are using Vicuna-13b-v1.5. We provide the hospital information 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". Both patients and hospital staff reported that our new LLM-based system has improved greatly, compared to our previous system (Gunson et al., 2022). In order to measure the improvement in its QA capabilities, we created a set of 100 in-domain questions that were designed to provide broad coverage of the modular system capabilities. These were a mix of hand-crafted and real questions asked by patients in our collected data. In-domain error rates, where incorrect or no information was given in response to the question, improved from 29.2% to 11.5%.

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 tips, teaching the user what the system can do (e.g. "I'm not sure, but I can help you with directions and menu information."). Many of these unexpected utterances can now be handled directly by the LLM.

#### 3.4.1 Grounding Responses to the Provided In-prompt Knowledge

Certain LLMs, like ChatGPT and Bard, are regularly asked general knowledge questions and expected to understand chit-chat. General LLM evaluation has therefore focused on their world knowledge learned at pre-training. For example, the popular Hugging Face Open LLM benchmark (the de facto standard leaderboard) ranks each model based on their performance across four tasks: (1) The AI2 Reasoning Challenge (Clark et al., 2018), a set of grade-school science questions; (2) MMLU (Hendrycks et al., 2020), a set of elementary level questions covering mathematics, US history, computer science, law, and more; (3) HelloSwag (Zellers et al., 2019), testing whether the model can select "what will happen next?" given a common sense scenario and some options; and (4) TruthfulQA (Lin et al., 2022), a set of 817 questions on various topics, like law and politics.

These corpora highlight the field's effort to reduce model hallucination. It is vital to clarify that they focus on hallucination reduction of outputs generated from the LLM's *static world knowledge*. In fact, this world knowledge can generate 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 before their appointment, and this response would result in a hospital patient going hungry. Other examples include how long a patient must wait for their medication to wear off before driving (Addlesee, 2024).

To tackle this problem, we must coax the LLM to ground its response to the in-prompt knowledge given at runtime, and not rely on non-domain-specific and potentially out-of-date knowledge learned at pre-training. To measure the impact of in-prompt grounding strategies, we used 50 questions from our project paired with a text passage. We do not always know what an LLM is trained on, and this could potentially include the website of our real hospital, so this passage described a fictitious hospital that no LLM could possibly know. We provide four prompts:

**Basic**: The passage followed by the question.

**Jodie**: Our prompt provides the passage as a quote by Jodie W. Jenkins, a fictitious non-celebrity name (according to Google). We then ask the LLM to answer according to Jodie. The exact pattern is this: 'Jodie W. Jenkins said "PASSAGE". Answer according to Jodie W. Jenkins. QUESTION'.

**Expert**: In order to ensure any prompt-grounding benefit is not simply a result of adding "according to", we again provide the passage as a quote by Jodie W. Jenkins, but add "Answer according to

Table 3: Knowledge grounding results. 🟩 indicates an improvement compared to the ‘basic’ prompt. 🟥 indicates a performance drop compared to the ‘basic’ prompt. **Bold** marks the best scores per model (Addlesee, 2024).

| LLM | Basic Prompt | | Jodie Prompt | | Expert Prompt | | Wikipedia Prompt | |
|---|---|---|---|---|---|---|---|---|
| | Quip | Acc | Quip | Acc | Quip | Acc | Quip | Acc |
| Dolly-12b | 38.71 | 36 | 35.74 | **42** | 28.08 | 32 | **39.21** | 34 |
| GPT-4 | 41.04 | 94 | **42.92** | **98** | 42.61 | 92 | 38.66 | 90 |
| Llama-7b-chat | 43.06 | 56 | **44.56** | **84** | 41.64 | 72 | 40.84 | 74 |
| Llama-13b-chat | **48.51** | **60** | 41.18 | **60** | 44.04 | 50 | 44.29 | 58 |
| Llama-70b-chat | 44.10 | 64 | **58.73** | **82** | 52.44 | 70 | 53.78 | 68 |
| Llama-70b-chat (0.95 temp) | 44.52 | 68 | **53.18** | **80** | 52.01 | 70 | 52.82 | 68 |
| Vicuna-13b-v1.1 | 64.93 | 46 | **80.95** | **54** | 29.17 | 12 | 31.93 | 26 |
| Vicuna-13b-v1.5 | 40.97 | 70 | **41.14** | **74** | 36.30 | 52 | 34.17 | 56 |

UnitedHealth” instead of Jodie W. Jenkins.

**Wikipedia**: The Expert prompt with one word replaced. The expert name is set to “Wikipedia”.

### 3.4.2 Response Grounding Results

In related work, Weller et al. (2023) measured LLM grounding to world knowledge. In order to measure how well an LLM’s output was grounded to Wikipedia, they devised a metric: QUIP-score. This score is the character n-gram precision of the generated output compared to the source corpus. It is a useful metric in our case too, as we can measure how precisely each LLM’s output is grounded in the given in-prompt knowledge. This focus on precision also punishes a model’s output when it hallucinates – our goal here too. Using our corpus (Addlesee, 2024), we used this QUIP-score and the answer’s accuracy to measure in-prompt grounding performance, as grounding is impractical if it does not preserve QA performance.

Table 3 illustrates the impressive performance of our ‘Jodie’ prompt. The Quip-score did decrease for two of the models, but the accuracy never deteriorated, and increased by up to 28% (mean: 10%). Even though the ‘Expert’ and ‘Wikipedia’ prompts differ from the ‘Jodie’ prompt by just one name, they generate more text that is not contained in the given prompt (as shown by the lower Quip-scores), and these additional hallucinations result in an accuracy drop.

Our current SDS utilises this ‘Jodie’ prompt in component (D) in Figure 2 to improve in-prompt grounding, reducing potential user harm.

## 3.5 Gesture Generation

As discussed in Section 1, MPCs are far more complex than two-party interactions. The SDS must track who said what to whom (Gu et al., 2022b), track the goals of multiple users (Addlesee et al., 2023d), and generate responses *to specific addressees*. As our SDS is embodied by the ARI social robot (Cooper et al., 2020), we can produce helpful gestures with its controllable arms, head, and eyes. While some gestures are charming, like facing the robot’s palms upward when welcoming a user to an interaction, other gestures are more functional. The robot can look at the user it is addressing, point when giving directions, and indicate that it is passing the turn to another user with its arm. These functional gestures are what we evaluate here. In component (E) in Figure 2, you can see that we generate gestures using the Vicuna-13b-v1.5 LLM (component (F)) in parallel with the grounded answer generation. In the prompt, we provide some examples of functional gestures, using the gesture tags that the robot expects (Cherakara et al., 2023). The answer text is passed to ARI’s text-to-speech, and the generated gesture tags are passed to ARI’s movement controls. We do not generate gestures when listening to the user, as the microphones become saturated by ego-noise (motor sounds), and the ASR fails to hear the user’s utterance (Addlesee et al., 2023b).

We annotated a set of 110 generated system responses with gold functional gesture tags. Using our gesture generation method, the generated gestures were accurate 86% of the time. Generating an incorrect gesture (e.g. pointing in the wrong direction) is more problematic than missing a gesture, and the gesture generation precision was 0.91.

## 4 Conclusions and Future Work

We have iteratively developed and deployed a multimodal, multi-party spoken dialogue system in a hospital memory clinic. This SDS is embodied by the ARI social robot, allowing us to generate gestures in addition to speech. 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 (improving accessibility for people with memory impairment), 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.

## Ethical Considerations

Some LLMs, like ChatGPT, can only be used through an API. This is a huge privacy concern, especially in the healthcare setting. Even if participants were instructed carefully, it is impossible to ensure they would not reveal personally identifiable information – this problem is exacerbated in a memory clinic setting (Addlesee and Albert, 2020). For this reason, we must use more open and transparent LLMs (Liesenfeld et al., 2023). We selected Vicuna-13b-v1.5 as it was the best performing model that could run on our hardware.

In Section 3.4 we detailed our in-prompt hallucination reduction efforts, but these will never reach zero. Hospital staff run the experiments, so they can correct the robot if it ever produces a hospital-related hallucination. This is also why we do not provide the SDS with any personal information like patient appointment schedules – we do not want to cause confusion.

In a real deployment, prompt poisoning could be an issue. A bad actor can manipulate the system to output incorrect responses through dialogue. This is not possible in our data collection, as we reset the system between participants (the patients are also unlikely to be bad actors). If deployed, speaker diarization and dialogue history deletion can mitigate this risk, but it is critical to highlight that LLMs can be manipulated.

## Acknowledgements

This research was funded by the EU H2020 program under grant agreement no. 871245 (`https://spring-h2020.eu/`). We would also like to thank our anonymous reviewers for their time and valuable feedback.

## References

Angus Addlesee. 2023. Voice assistant accessibility. In *Proceedings of the 13th International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

Angus Addlesee. 2024. *Incremental Multi-party Conversational AI for People with Dementia*. Ph.D. thesis, Heriot-Watt University.

Angus Addlesee and Pierre Albert. 2020. Ethically collecting multi-modal spontaneous conversations with people that have cognitive impairments. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 15.

Angus Addlesee and Marco Damonte. 2023a. Understanding and answering incomplete questions. In *Proceedings of the 5th Conference on Conversational User Interfaces*.

Angus Addlesee and Marco Damonte. 2023b. Understanding disrupted sentences using underspecified abstract meaning representation. In *Proceedings of INTERSPEECH 2023*, pages 1224–1228.

Angus Addlesee, Daniel Denley, Andy Edmondson, Nancie Gunson, Daniel Hernández Garcia, Alexandre Kha, Oliver Lemon, James Ndubuisi, Neil O'Reilly, Lia Perochaud, Raphaël Valeri, and Miebaka Worika. 2023a. Detecting agreement in multi-party dialogue: evaluating speaker diarisation versus a procedural baseline to enhance user engagement. In *Proceedings of the workshop on advancing GROup UNderstanding and robots aDaptive behaviour (GROUND)*.

Angus Addlesee, Arash Eshghi, and Ioannis Konstas. 2019. Current challenges in spoken dialogue systems and why they are critical for those living with dementia. In *Dialogue for Good (DiGo)*.

Angus Addlesee, Ioannis Papaioannou, and Oliver Lemon. 2023b. Building for speech: Designing the next generation of social robots for audio interaction. In *Proceedings of the 2nd Workshop on Working with Trouble and Failures in Conversation Between Humans and Robots (WTF)*.

Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernández Garcia, Christian Dondrup, and Oliver Lemon. 2023c. Data collection for multi-party task-based dialogue in social robotics. In *Proceedings of the 13th International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernández García, Christian Dondrup, and Oliver Lemon. 2023d. 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*.

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*, pages 114–130. Springer.

Apple. 2022. Use accessibility features with siri on iphone. [Online; accessed 14-April-2023].

Peter Auer. 2018. Gaze, addressee selection and turn-taking in three-party interaction. *Eye-tracking in interaction: Studies on the role of eye gaze in dialogue*, pages 197–231.

Janet Beavin Bavelas and Jennifer Gerwing. 2011. The listener as addressee in face-to-face dialogue. *International Journal of Listening*, 25(3):178–198.

Veronica Boschi, Eleonora Catricala, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F Cappa. 2017. Connected speech in neurodegenerative language disorders: a review. *Frontiers in psychology*, 8:269.

Neeraj Cherakara, Finny Varghese, Sheena Shabana, Nivan Nelson, Abhiram Karukayil, Rohith Kulothungan, Mohammed Afil Farhan, Birthe Nesset, Meriam Moujahid, Tanvi Dinkar, Verena Rieser, and Oliver Lemon. 2023. FurChat: An embodied conversational agent using LLMs, combining open and closed-domain dialogue with facial expressions. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 588–592, Prague, Czechia. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

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*, pages 175–182, Prague, Czechia. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

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*, pages 745–751.

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 Proceedings*.

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

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*.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*.

Press Furhat Robotics. 2015. Franny, frankfurt airport's new multilingual robot concierge can help you in over 35 languages. *Furhat Robotics Press Release*.

Charles. Goodwin. 1981. *Conversational organization: interaction between speakers and hearers*. Academic Press.

Jia-Chen Gu, Chao-Hong Tan, Chongyang Tao, Zhen-Hua Ling, Huang Hu, Xiubo Geng, and Daxin Jiang. 2022a. HeterMPC: A Heterogeneous Graph Neural Network for Response Generation in Multi-Party Conversations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5086–5097.

Jia-Chen Gu, Chongyang Tao, and Zhen-Hua Ling. 2022b. 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)*.

Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. MPC-BERT: A pre-trained language model for multi-party conversation understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 3682–3692.

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)*, pages 1352–1357. IEEE.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Christine Howes, Ptarick GT Healey, Matthew Purver, and Arash Eshghi. 2012. Finishing each other's... responding to incomplete contributions in dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. GSN: A graph-structured network for multi-party dialogues. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*.

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*, pages 143–152.

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):1–32.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, pages 1–6.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Meriam Moujahid, Helen Hastie, and Oliver Lemon. 2022. Multi-party interaction with a robot receptionist. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 927–931. IEEE.

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*, pages 120–123.

Laura Panfili, Steve Duman, Andrew Nave, Katherine Phelps Ridgeway, Nathan Eversole, and Ruhi Sarikaya. 2021. Human-ai interactions through a gricean lens. *Proceedings of the Linguistic Society of America*, 6(1):288–302.

Ioannis Papaioannou, Amanda Cercas Curry, Jose L Part, Igor Shalyminov, Xinnuo Xu, Yanchao Yu, Ondrej Dušek, Verena Rieser, and Oliver Lemon. 2017. Alana: Social dialogue using an ensemble model and a ranker trained on user feedback. *Alexa Prize Proceedings*.

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*, pages 1–12.

Laura Schauer, Jason Sweeny, Charlie Lyttle, Zein Said, Aron Szeles, Cale Clark, Katie McAskill, Xander Wickham, Tom Byars, Daniel Hernández Garcia, Nancie Gunson, Angus Addlesee, and Oliver Lemon. 2023. Detecting agreement in multi-party conversational ai. In *Proceedings of the workshop on advancing GROup UNderstanding and robots aDaptive behaviour (GROUND)*.

Antoine Slegers, Renee-Pier Filiou, Maxime Montembeault, and Simona Maria Brambati. 2018. Connected speech features from picture description in alzheimer's disease: A systematic review. *Journal of Alzheimer's Disease*, 65(2):519–542.

Francesco Tonini, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. 2023. Object-aware gaze target detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21860–21869.

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*, pages 201–211. Springer.

Evgenios Vlachos, Anne Faber Hansen, and Jakob Povl Holck. 2020. A robot in the library. In *International conference on human-computer interaction*, pages 312–322. Springer.

Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2023. " according to..." prompting language models improves quoting from pre-training data. *arXiv preprint arXiv:2305.13252*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

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*, volume 36, pages 11765–11773.