Trust in Autonomous Human–Robot Interaction
An In-Person Pilot Study

M.C. Lau[a,1,*], Shauna Heron[a,2]

[a] *Laurentian University, Bharti School of Engineering,*

[b] *Laurentian University, School of Social Sciences,*

Abstract

This study implements a multi-stage collaborative task system where participants collaborate with the Misty-II social robot to solve a who-dunnit type task. The system utilizes an autonomous, mixed-initiative dialogue architecture with affect-responsive capabilities.

Keywords: keyword1, keyword2

! TODO

Manually score each dialogue series.
For each interaction and stage:

- did the participant ask for help?
- how many times?
- did the robot give useful help?
- did the robot give misleading or incorrect help?
- did the robot stick to the policy?
- how many times did the robot fail to understand the participant?

For each task:

- is there evidence that the robot helped complete the task?
- is there evidence that the participant solved the problem without help?

From our scoring, we make a variable binary for participants who had

*Corresponding author

Email addresses: mclau@laurentian.ca (M.C. Lau), sheron@laurentian.ca (Shauna Heron)

[1] This is the first author footnote.

[2] Another author footnote, this is a very long footnote and it should be a really long footnote. But this footnote is not yet sufficiently long enough to make two lines of footnote text.

serious communication breakdowns/issues defined by abandonding tasks altogether, and inability for STT to translate participant speech to text. These participants will be excluded from the final analysis.

Human–robot collaboration (HRC) has become a central topic across engineering, computer science, and the social sciences as robots increasingly move from controlled laboratory settings into everyday collaborative roles. In many emerging applications, collaboration depends not only on physical coordination but also on shared problem-solving through dialogue, where robots must reason, communicate, and adapt in real time. Understanding how humans perceive and respond to such systems is therefore critical for designing robots that can function as effective collaborators rather than passive tools.

A key factor shaping successful collaboration in human–robot interaction (HRI) is trust. Trust influences whether users are willing to rely on robotic systems, accept their guidance, and remain engaged during joint tasks, particularly in situations characterized by uncertainty or incomplete information. Prior work has shown that trust affects both subjective perceptions—such as perceived reliability or intent—and objective outcomes including task performance, compliance, and cooperation. As a result, a substantial body of research has focused on measuring trust in HRI, leading to the development of standardized instruments for assessing users' evaluations of robot behaviour across industrial, medical, and social contexts.

Despite this progress, much of the existing literature on trust in HRI is based on interactions conducted under highly controlled or simulated conditions. In many studies, robot behaviour is scripted, partially simulated, or mediated through Wizard-of-Oz paradigms, where a human operator covertly controls aspects of the robot's behaviour. While these approaches are valuable for isolating specific design factors and testing early hypotheses, they also mask many of the failures and inconsistencies that characterize autonomous systems in real-world use. Speech recognition errors, delayed or inappropriate responses, misinterpretations of user intent, and limitations of affect sensing are not peripheral issues but central features of deployed autonomous robots. These imperfections are likely to play a decisive role in shaping trust, yet they remain underexplored in empirical HRI research.

The present pilot study addresses this gap by examining trust and collaboration in an in-person interaction with a fully autonomous social robot operating within predefined behavioural constraints. Using a between-subjects design, participants collaborated with a robot during an immersive, dialogue-driven puzzle game in which the robot acted as a diegetic game guide and partner. The task required shared problem-solving through conversation, with participants seeking hints, advice, and support from the robot while navigating the game environment. Crucially, all interaction management—including speech-based dialogue, task progression, and affect-responsive behaviour—was handled autonomously by the

robot without human intervention.

Two versions of the robot were compared. In one condition, the robot was designed to be proactive and responsive, adapting its behaviour based on participant affect and conversational cues. In the other condition, the robot provided assistance only when explicitly requested, offering a more reactive interaction style. This manipulation allowed us to examine reminder differences in autonomy and responsiveness influence trust perceptions and collaborative performance under otherwise identical task demands.

To support this interaction, we developed an autonomous spoken-language system integrated with automatic speech recognition and affect detection on the Misty-II robot platform. The system we developed enables the robot to recognize speech, manage dialogue state, maintain conversational context, and generate coordinated verbal responses alongside facial expressions and head and arm movements. Rather than optimizing for flawless performance, the system was designed to reflect realistic capabilities and limitations of contemporary social robots.

By combining post-interaction trust measures with behavioural and task-level outcomes, this study aims to contribute empirical evidence on how trust is shaped in fully autonomous HRI scenarios. The focus is not on demonstrating idealized interaction under perfect conditions, but on examining trust as it emerges through realistic human–robot collaboration, where uncertainty, interactional breakdowns, and adaptive behaviour are unavoidable. In doing so, this work seeks to inform the design and evaluation of affect-responsive autonomous robots intended for real-world collaborative settings.

0.1. Hypotheses

The primary objective of this study was to examine how differences in robot interaction policy influence trust and collaboration during fully autonomous, in-person human–robot interaction. Based on prior work linking robot responsiveness, affective behavior, and trust in HRI, we formulated the following hypotheses.

H1: Participants interacting with a responsive, affect-adaptive robot will report higher post-interaction trust than participants interacting with a neutral, reactive robot.

H2: Participants in the responsive condition will demonstrate greater engagement with the robot during the collaborative tasks, reflected in increased voluntary interaction and reliance on robot input during problem solving.

H3: Differences in trust and engagement will be most pronounced during the open-ended collaborative task, where assistance from the robot is optional rather than required.

0.2. Methods

0.2.1. Sample and recruitment

Participants (n = 29) were recruited from the Laurentian University community through word of mouth and the SONA participant recruitment system. Eligibility criteria required participants to be adults (18 years or older), fluent in English, with normal or corrected-to-normal hearing and vision, and no prior experience interacting with the Misty-II robot. Participants received a $15.00 gift card as compensation for their time. All procedures were approved by the university's Research Ethics Board. The Misty-II robot used in this study was purchased through grant funding from the IAMGOLD President's Innovation Fund. Sample characteristics are summarized in Table 1.

0.2.2. Experimental design

The study employed a between-subjects design with robot interaction policy as the sole experimental factor. Participants interacted with the Misty-II robot in a shared physical workspace that included both the robot and a participant-facing computer interface. The interface was used to present brief task instructions, collect participant inputs, and manage transitions between task stages. Critically, the interface did not serve as a control mechanism for the robot. Instead, the robot autonomously monitored task state and participant inputs via the interface and adapted its dialogue and behavior accordingly, without any real-time human intervention (see Figure 1).

Participants collaborated with the robot during an immersive puzzle game in which the robot functioned as a diegetic game guide and collaborative partner. The interaction was fully autonomous in both conditions, and both versions of the robot were subject to the same sensory and interaction constraints inherent to real-world operation, including speech recognition variability and response timing delays. The only manipulation between conditions was the robot's interaction policy.

Participants were randomly assigned to one of two conditions:

RESPONSIVE (experimental): The robot adopted a warm, emotionally engaged, and proactive interaction style, adapting its responses based on detected participant affect, dialogue context, and task demands.

CONTROL (baseline): The robot followed a neutral, reactive interaction policy, providing information and assistance only when explicitly requested, without affect-responsive adaptation.

0.2.3. Task structure

The game consisted of five sequential stages designed to elicit interaction under differing collaboration and dependency conditions, following established approaches in HRI task design (Lin et al., 2022). Total session duration was approximately 15 minutes.

Figure 1: Experimental setup showing the autonomous robot and participant-facing task interface used during in-person sessions. Participants entered task responses and navigated between task stages using the interface, while the robot autonomously tracked task state and adapted its interaction based on participant input. No real-time human intervention occurred during the interaction.

Stage 1: Greeting. The robot introduced itself and engaged in brief rapport-building interaction.

Stage 2: Mission brief. The robot explained the narrative context and overall objectives of the task.

Stage 3: Task 1 (robot-dependent reasoning). Participants completed a constrained "who-dunnit" task.

Stage 4: Task 2 (open-ended collaborative problem solving). Participants worked to determine the location of the missing robot using technical logs.

Stage 5: Wrap-up. The robot provided feedback and concluded the interaction.

0.2.4. Task 1: Robot-dependent collaborative reasoning

In the first task, participants were required to identify a suspect from a 6×4 grid of 24 candidates by asking the robot a series of yes/no questions about the suspect's features (e.g., hair color, accessories, clothing). The grid was displayed on the interface, while questions were posed verbally to the robot. The robot possessed the ground-truth information necessary to evaluate each question and provide correct responses.

Successful completion of this task was therefore dependent on interaction with the robot, creating a forced collaborative dynamic in which the robot served as an essential informational partner. Participants were required to coordinate questioning strategies with the robot to narrow down the correct suspect within a five-minute time limit. The structured nature of the task ensured consistent interaction demands across participants and conditions.

0.2.5. Task 2: Open-ended problem solving with advisory robot support

The second task involved a more open-ended problem-solving scenario. Participants were presented with multiple technical logs through a simulated terminal interface and were asked to determine the location of the missing robot. Unlike Task 1, the robot did not have access to ground-truth information or the contents of the logs. The robot's assistance was limited to general problem-solving support derived from its language model, such as explaining how to interpret logs, suggesting reasoning strategies, or prompting participants to reflect on inconsistencies.

Participants could complete this task independently or choose to solicit assistance from the robot. The robot could ask clarifying questions about what the participant observed in the logs, and participants could likewise ask the robot for guidance. This design positioned the robot as a collaborative reasoning partner rather than an authoritative source and allowed collaboration to emerge voluntarily rather than being enforced by task structure (Lin et al., 2022).

0.2.6. Wrap-up and debrief

After all responses were submitted, correct answers were displayed to participants. During the wrap-up stage, the robot engaged in a brief debriefing

interaction, acknowledging task outcomes and thanking participants for their involvement before prompting them to report back to the researcher.

0.2.7. In-person procedure

Participants completed a pre-interaction questionnaire administered via Qualtrics prior to their in-person session. This questionnaire included informed consent, demographic information, the Negative Attitudes Toward Robots Scale, and a measure of Need for Cognition. Balanced random assignment was also completed in this step (need mention no-shows that threw off the balance of the group assignments). Due to variability in timing between pre-interaction questionnaires and in-person sessions, these measures were treated as baseline covariates rather than formal pre-test measures.

At the start of the in-person session, participants were seated in front of the Misty-II robot and instructed to begin the interaction by clicking a start button on the interface. They were given brief guidance on effective communication with the robot, including waiting for a visual indicator on the robot before speaking. Once participants indicated readiness, the researcher left the room and closed the door, leaving the participant and robot to complete the tasks without human presence.

Following task completion, participants exited the room and completed a post-interaction survey assessing trust using the Trust Perception–HRI scale and the Trust in Industrial Human–Robot Collaboration scale. Participants then engaged in a written and verbal debrief with the researcher. Participants were informed that they could terminate the session at any time without penalty. All participants completed the full procedure, with total session duration averaging approximately 30 minutes, and received compensation upon completion.

1. Results

1.1. Participant characteristics and baseline measures

Participants in the control and responsive conditions were comparable with respect to pre-interaction demographic characteristics, academic background, prior experience with robots, and baseline attitudes toward robots. Importantly, Negative Attitudes Towards Robots (NARS) and Need for Cognition scores were similar across groups, indicating that post-interaction differences are unlikely to reflect pre-existing attitudes (see Table 1).

1.2. Post-Interaction Trust Differences

Descriptive comparisons of participant-level post-test scores indicated an approximately 12 point difference in post-test Trust Perception Scale-HRI scores (M 75 vs M 63) and a 27 point difference in the Trust in Industrial Human-robot Collaboration scale (M 39 vs M 66) between conditions, although in the first scale differences did not reach conventional significance under a two-sample t-test (p=.10), the second scale was significantly different between groups (p=.007).

To test these findings further, we fitted several Bayesian hierarchical models were fitted (estimated using MCMC sampling with 4 chains of 4000 iterations and a warmup of 1000) to predict Robot HRI-trust and Trust in HRI Collaboration by experimental group (formula: robot_trust_post ~ group). The model included session_id and trust_items as random effects (formula: list(~1 | session_id, ~1 | trust_items)). Both models indicated higher post-interaction trust scores in the responsive robot condition across both trust-related scales (posterior median differences 8-15 points on a 0–100 scale).

For the Trust in Industrial Robots outcome, the responsive condition showed a robust positive effect on post-task trust ratings. The estimated group difference was 14.5 points (95% CrI [5.62, 23.22]), exceeding between-session variability and remaining stable after accounting for item-level effects, with a 95% chance of being large (>6.94). In contrast, for the HRI trust perception scale, the estimated group effect was smaller at ~ 9 points and more uncertain 95% CrI [-1.72, 19.25], with 65% chance of being large (>6.87).

The posterior probability that the responsive condition increased trust was greater than 95% for both measures, suggesting a robust directional effect despite substantial individual variability. Sensitivity analyses using substantially wider priors yielded nearly identical posterior estimates for the group effect, indicating that results were not driven by prior specification.

In addition to directional effects, the posterior probability that the responsive condition increased trust by at least five points was 77% in HRI-trust perception and 98% in the Collaborative Trust in Industrial Robots scale, suggesting a reasonable likelihood of a practically meaningful effect. Moreover, in the latter collaboration scale, there is an 85% likelihood of an effect-size greater than 10 points.

1.3. Trust subscale patterns

1.4. Interaction dynamics and task performance

1.4.1. Task performance

Objective task accuracy did not differ between conditions across any task-level measures except suspect accuracy (robot dependendant task), indicating that increased trust was only attributable to improved task success when interaction was necessary to complete accurately.

Despite similar task accuracy, interactions in the responsive condition were characterized by longer durations, slower response times, and a higher number of AI-detected engaged responses. These findings suggest that responsiveness altered the interaction dynamics and affective tone rather than task outcomes.

1.5. Individual differences and correlational patterns

As expected, we found that higher Need for Cognition (NFC) scores were negatively associated with Negative Attitudes Towards Robots (NARS), indicating that individuals who enjoy effortful thinking tend to have more positive attitudes

Characteristic	**N**	**CONTROL** N = 8[1]	**RESPONSIVE** N = 14[1]	**p-value**[2]
post_trust	22	39 (22)	66 (21)	**0.007**
post_trust_reliability	22	42 (27)	66 (17)	**0.029**
post_trust_perception	22	36 (24)	56 (24)	0.078
post_trust_feelings	22	47 (31)	76 (25)	**0.032**
Post-Task Trust Perception	22	63 (16)	75 (19)	0.10
Suspect ID Accuracy	22	3 / 8 (38%)	9 / 14 (64%)	0.38
Status Accuracy	22	6 / 8 (75%)	9 / 14 (64%)	>0.99
building_correct	22	6 / 8 (75%)	12 / 14 (86%)	0.60
floor_correct	22	6 / 8 (75%)	12 / 14 (86%)	0.60
zone_correct	22	4 / 8 (50%)	4 / 14 (29%)	0.39
Total Task Accuracy	22	3.13 (1.13)	3.29 (1.14)	0.70
Overall Task Accuracy	22	0.63 (0.23)	0.66 (0.23)	0.70
Dialogue Turns	22	38 (6)	37 (11)	0.19
Avg Task Duration (mins)	22	14.36 (2.18)	17.26 (6.67)	0.13
Avg Response Time (ms)	22	13.19 (0.90)	17.57 (2.30)	**<0.001**
Silent Periods	22	5.75 (2.19)	5.14 (2.98)	0.39
Engaged Responses	22	2.50 (2.20)	3.71 (1.82)	0.089
Frustrated Responses	22	0.63 (0.74)	0.86 (1.23)	0.94
n_neg	22	0.88 (0.83)	0.93 (1.21)	0.80

[1]Mean (SD); n / N (%)

[2]Wilcoxon rank sum test; Wilcoxon rank sum exact test; Fisher's exact test

towards robots. This relationship is consistent with prior literature suggesting that cognitive engagement is associated with openness to new technologies. In terms of NARS subscales, NFC was negatively correlated with all three subscales, but significantly so only in the domain of Situations of Interaction with Robots. This suggests that individuals with higher NFC are less likely to hold negative attitudes across various dimensions of robot interaction but especially around direct interaction with robots.

–> how to talk about post-interaction correlations w/pre-interaction measures Several behavioural and task-level measures were correlated with post-interaction trust, consistent with the interpretation that trust judgments were shaped by interaction quality; these variables were not included as covariates in primary models to avoid conditioning on potential mediators.

Baseline negative attitudes toward robots were negatively correlated with post-interaction trust, with the strongest associations observed for affective trust subscales. In contrast, objective task performance was selectively associated with perceived reliability. Need for cognition was negatively correlated with negative robot attitudes and interaction-level negative affect, suggesting that individual differences contributed to variability in trust responses.

1.5.1. Model robustness and predictive checks

Sensitivity analyses using alternative prior specifications yielded substantively similar estimates, and leave-one-out cross-validation indicated comparable predictive performance between models with and without the group effect.

! TO DO:

- add subscale column to long format data
- run an analysis of performance by robot-dependent versus robot-independent tasks
- write up a future directions section for the planned larger study
- talk about unexpected language issues with people signing up with difficultly speaking and understanding english which cuased problems with asr and interaction
- run analysis of dialogue dynamics included Bertopic or some other analysis of the actual content of the conversations/interactions

2. Discussion

Mention language confounders!!

The second task was intentionally designed to be sufficiently challenging that completing it within the allotted time was difficult without assistance. This ensured that interaction with the robot represented a meaningful opportunity for collaboration rather than a trivial or purely optional exchange. By contrasting a

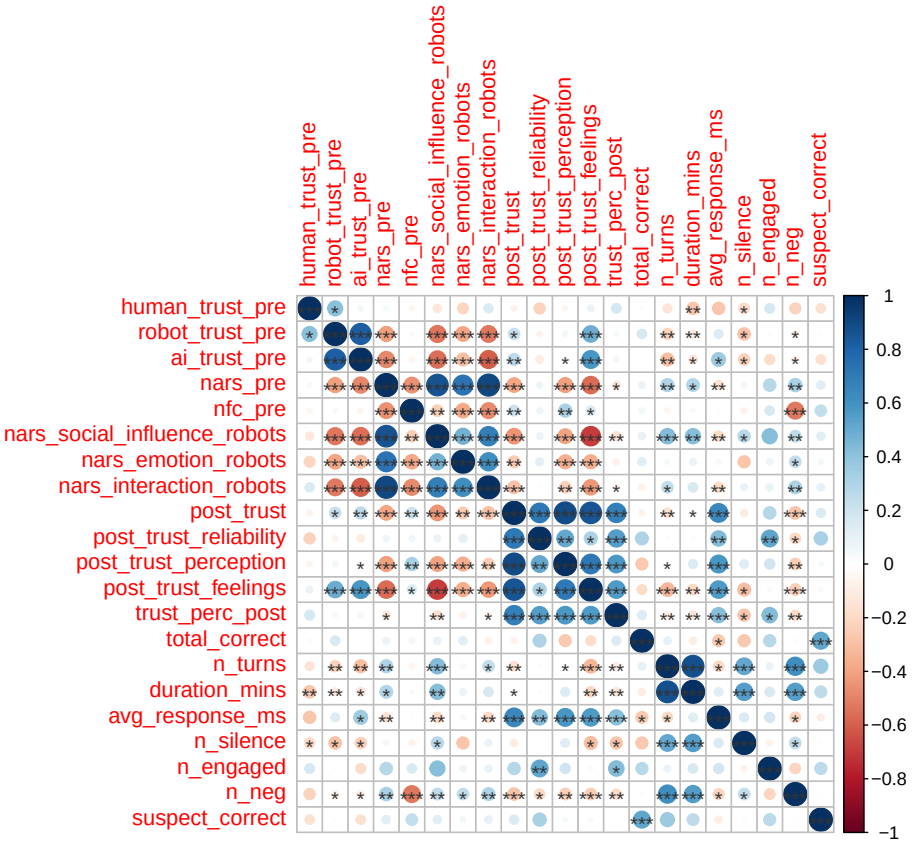
human_trust_pre
robot_trust_pre
ai_trust_pre
nars_pre
nfc_pre
nars_social_influence_robots
nars_emotion_robots
nars_interaction_robots
post_trust
post_trust_reliability
post_trust_perception
post_trust_feelings
trust_perc_post
total_correct
n_turns
duration_mins
avg_response_ms
n_silence
n_engaged
n_neg
suspect_correct
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Figure 2

robot-dependent task with an open-ended advisory task, the study examined trust formation across interaction contexts that varied in both informational asymmetry and reliance on the robot.

This pilot study examined trust outcomes following in-person interaction with an autonomous social robot under two interaction policies: a responsive, affect-adaptive condition and a neutral, non-responsive control condition. By leveraging a fully autonomous dialogue system integrated with speech recognition and affect detection, the study aimed to evaluate how robot responsiveness influences trust formation in realistic human–robot collaboration scenarios.

Descriptive comparisons of post-interaction measures indicated that participants in the responsive condition reported consistently higher trust across all trust measures, with differences ranging from approximately 8 to 16 points on a 0–100 scale, although uncertainty remained high given the small sample. Notably, the responsive condition did not differ from control in objective task accuracy, suggesting that increased trust was not driven by improved task success. Instead, responsive interactions were characterized by longer durations, slower response times, and a higher number of AI-detected engaged responses, indicating a shift in interaction dynamics rather than performance.

Baseline negative attitudes toward robots were most strongly associated with affective components of trust rather than perceptions of reliability, suggesting that pre-existing attitudes primarily shape emotional responses to interaction rather than judgments of system competence. Conversely, objective task performance was selectively associated with perceived reliability, indicating that participants distinguished between affective and functional aspects of trust.

Future work with larger samples could formally test mediation pathways linking robot responsiveness, interaction fluency, affective responses, and trust judgments, as well as moderation by baseline attitudes toward robots and need for cognition.

Participants in the responsive condition also exhibited higher levels of AI-detected engagement during interaction, as indexed by a greater number of responses classified as positive affect (t-test result). This suggests that responsive behaviours altered the affective tone of the interaction itself.

3. Appendix

3.1. Dialogue Coding Scheme

3.1.1. Task Outcome Layer (Stage-Level)

Table 1

Characteristic	N	CONTROL N = 8[1]	RESPONSIVE N = 14[1]	p-value[2]
Gender	21			0.66
Woman		3 / 8 (38%)	7 / 13 (54%)	
Man		5 / 8 (63%)	6 / 13 (46%)	
Age Group	21			0.37
18-24		4 / 8 (50%)	6 / 13 (46%)	
25-34		2 / 8 (25%)	2 / 13 (15%)	
34-44		0 / 8 (0%)	4 / 13 (31%)	
45+		2 / 8 (25%)	1 / 13 (7.7%)	
Program	19			0.94
Psychology		1 / 8 (13%)	1 / 11 (9.1%)	
Engineering		2 / 8 (25%)	1 / 11 (9.1%)	
Computer Science		3 / 8 (38%)	5 / 11 (45%)	
Earth Sciences		0 / 8 (0%)	1 / 11 (9.1%)	
Other		2 / 8 (25%)	3 / 11 (27%)	
Experience w/Robots	22	5 / 8 (63%)	3 / 14 (21%)	0.081
Native English Speaker	22			>0.99
Native English		4 / 8 (50%)	8 / 14 (57%)	
Non-Native English		4 / 8 (50%)	6 / 14 (43%)	
NARS Overall	22	37 (10)	38 (7)	0.89
Need for Cognition	22	3.92 (0.74)	3.77 (0.78)	0.81

[1]n / N (%); Mean (SD)
[2]Fisher's exact test; Wilcoxon rank sum test

Variable	Type	Description
`task_outcome`	categorical	Final task status (`completed`, `timeout`, `skipped`, `partial`, `abandoned`). Exactly one per task.
`task_completed`	binary	Task goal was fully completed within the allotted time.
`task_timed_out`	binary	Task ended due to expiration of the time limit before completion.
`task_skipped`	binary	Participant explicitly skipped or advanced past the task without completing it.
`task_partially_completed`	binary	Task progress was made, but the full solution was not reached.
`task_abandoned`	binary	Participant disengaged or stopped attempting the task before timeout.
`task_time_remaining_sec`	numeric	Time remaining (in seconds) when the task ended; 0 if timed out.
`task_completed_without_help`	binary	Task was completed without any help requests to the robot.
`task_required_robot_help`	binary	At least one robot help interaction was required for task completion.

3.1.2. Dialogue Interaction Layer (Turn-Level)

3.1.2.1. Human Turn Codes.

Variable	Type	Description
`human_help_request`	binary	Participant explicitly or implicitly asks the robot for help or guidance.
`human_reasoning_self`	binary	Participant articulates their own reasoning or problem-solving independent of the robot.

Variable	Type	Description
human_confusion	binary	Participant expresses confusion or uncertainty.
human_confirmation_seeking	binary	Participant seeks confirmation of a tentative belief or solution.
human_ignores_robot	binary	Participant proceeds without engaging with the robot's prior input.

3.1.2.2. Robot Turn Codes.

Variable	Type	Description
robot_helpful_guidance	binary	Robot provides accurate, task-relevant guidance.
robot_misleading_guidance	binary	Robot provides misleading or incorrect guidance.
robot_factually_incorrect	binary	Robot states information that is objectively incorrect.
robot_policy_violation	binary	Robot violates stated system or task constraints.
robot_on_policy_unhelpful	binary	Robot adheres to policy but provides vague or non-actionable assistance.
robot_stt_failure	binary	Robot response reflects a speech-to-text or input understanding failure.
robot_clarification_request	binary	Robot asks the participant to repeat or clarify their input.

3.1.3. Affective Interaction Layer (Turn-Level)
3.1.3.1. Robot Affective Behavior.

Variable	Type	Description
`robot_empathy_expression`	binary	Robot expresses empathy, encouragement, or reassurance.
`robot_emotion_acknowledgement`	binary	Robot explicitly references an inferred participant emotional state.
`robot_affect_task_aligned`	binary	Robot's affective response is appropriate and supportive in context.
`robot_affect_misaligned`	binary	Robot's affective response is mistimed or disruptive to the task.

3.1.3.2. Human Affective Response.

Variable	Type	Description
`human_affective_engagement`	binary	Participant responds in a socially warm or emotionally engaged manner.
`human_social_reciprocity`	binary	Participant mirrors or responds to the robot's affective expression.
`human_anthropomorphic_language`	binary	Participant treats the robot as a social agent.
`human_emotional_disengagement`	binary	Participant responds in a curt, dismissive, or withdrawn manner.

3.1.4. Notes

- Turn-level variables are coded per dialogue turn.
- Task outcome variables are coded once per `session_id × stage`.
- Raw dialogue text was retained during coding and removed prior to aggregation.
- Multiple turn-level codes may co-occur unless otherwise specified.

3.2. Key Components of the System

This study implemented a multi-stage collaborative task system where participants collaborate with the Misty II social robot to solve a who-dunnit type task.

The system utilizes an autonomous, mixed-initiative dialogue architecture via langchain with affect-responsive capabilities.

1. Misty-II Robot: A programmable robot platform equipped with sensors and actuators for interaction.
2. Automated Speech Recognition (ASR): A speech-to-speech pipeline that processes spoken input from users and converts it into text for LLM processing then back to speech for output on the robot.
    - STT: Deepgram API for real-time speech-to-text conversion.
    - DistilRoBERTa-base fine-tuned on emotion classification for emotion detection from user utterances
    - LLM: Gemini API for processing text input and generating contextually relevant responses in JSON format
    - TTS: Misty-II text-to-speech (TTS) engine on 820 processor.
3. Langchain Dialogue Management: A system that manages the flow of conversation, ensuring coherent and contextually appropriate dialogue within a two-part collaborative task.
4. Collaborative-Tasks
    - Task 1: Whodunnit style task where human and robot collaborate to find a missing robot via the human asking Yes/No questions (process of elimination in 6x4 suspect grid) to the robot. Robot knows ground truth but can only answer Yes/No questions about suspect features. Can not directly describe the suspect or name them. (human can choose a random suspect to solve on their own but only 1 in 24 chance of being correct without robot help)
    - Task 2: Where is Atlas? Robot collaborates with human to find Atlas by deciphering cryptic system and sensor logs. Robot does not know the answer here and can only guide the human usinng its expertise and knowledge of computer systems and basic logical reasoning. (human can solve on their own but very difficult without robot help depending on participants technical background).
5. Flask-gui dashboard interface: A web-based interface/dashboard that allowed participants to interact with the tasks, view task-related information and input their answers to the questions. Responses were sent to the robot to signal task progression.
    - Task 1 dashboard: Displays the suspect grid and allows the user to select suspects and view their features.
    - Task 2 dashboard: Displays system logs and allows the user to input their findings.
6. Pre and post tests:
    - PRE-TESTS: Need for Cognition Scale (short); Negative Attitudes to Robots Scale (NARS);
    - POST-TESTS: Trust Perception Scale-HRI; 9 custom questions adapted from Charalambous et al. (2020) on trust in industrial human-robot collaboration;

4. Technical Specifications

4.1. System Overview

This study implements a multi-stage collaborative task system where participants collaborate with the Misty II social robot to solve a who-dunniti type task. The system utilizes an autonomous, mixed-initiative dialogue architecture with affect-responsive capabilities.

4.2. Hardware Platform

Robot: Misty II Social Robot (Furhat Robotics)

- Mobile social robot platform with expressive display, arm actuators, and head movement
- RGB LED for state indication
- RTSP video streaming (1920×1080, 30fps) for audio capture
- Custom action scripting for synchronized multimodal expressions

4.3. Software Architecture

4.3.1. Core System Components

Programming Language: Python 3.10

Primary Dependencies:

- `misty-sdk` (Python SDK for Misty Robotics API) - Robot control and sensor access
- `deepgram-sdk` (4.8.1) - Speech-to-text processing
- `ffmpeg-python` (0.2.0) - Audio stream processing
- `flask` (3.1.2) + `flask-socketio` (5.5.1) - Web interface for task presentation
- `duckdb` (1.4.0) - Experimental data logging database

4.3.2. Large Language Models

LLM Provider:

Google Gemini:

- Model: `gemini-2.5-flash-lite` (configurable via environment variable)
- Integration: `langchain-google-genai` with `google-generativeai` API
- Response format: JSON-only output (`response_mime_type: "application/json"`). This format is required by Misty-II for reliable parsing and for action execution.

LLM Configuration:

- Temperature: 0.7 (for balanced creativity and coherence)
- Memory: Conversation buffer memory with file-based persistence (`langchain.memory.ConversationBufferMemory`)
- Context window: Full conversation history maintained across interaction stages but reset between sessions.

4.4. LangChain Framework Integration

4.4.1. Core LangChain Components

Framework Version: `langchain-core` with modular provider packages

- `langchain` (meta-package)
- `langchain-community` (0.3.31)
- `langchain-google-genai` Gemini integration

4.4.2. ConversationChain Architecture

Memory Management (`ConversationChain` class in `conversation_chain.py`):

1. **Conversation Buffer Memory**:
   - Implementation: `langchain.memory.ConversationBufferMemory`
   - Storage: File-based persistent chat history (`FileChatMessageHistory`)
   - Format: JSON files in `.memory/` directory, one per participant session
   - Memory key: `"history"`
   - Return format: Message objects (full conversation context)
2. **Memory Reset Policy**:
   - Default: Reset on each new session launch
   - Archive previous session: Timestamped archive files stored in `.memory/archive/`
   - Configuration: `RESET_MEMORY` and `ARCHIVE_MEMORY` environment variables

4.4.3. Prompt Construction

Message Structure

(LangChain message types): `python [SystemMessage, *history_messages, HumanMessage]`

System Message Assembly:

- Core instructions (task framing, role definition)
- Personality instructions (mode-specific behaviour)
- Stage-specific instructions (current task context)
- Output format constraints (JSON schema specification)

```
Human Message Format:   {
"user": "<transcribed_speech>",
"stage": "<current_stage>",
"detected_emotion": "<emotion_label>",
"frustration_note": "<optional_alert>",
"timer_expired": "<task_id>",     ...   }
```

- JSON-encoded context variables passed alongside user input
- Enables LLM to access environmental state without breaking message history

4.4.4. Memory Persistence:

- Save after each turn: memory.save_context({"input": user_text}, {"output": llm_response})
- Maintains conversational coherence across multi-stage interaction
- Enables LLM to reference previous exchanges (e.g., "As I mentioned earlier…")

4.4.5. LangChain Design Rationale

Why LangChain for this application:

1. Memory abstraction: Automatic conversation history management without manual message list handling
2. Provider flexibility: Easy switching between Gemini and OpenAI without rewriting prompt logic
3. Message typing: Structured SystemMessage/HumanMessage/AIMessage types maintain role clarity
4. File persistence: Built-in FileChatMessageHistory enables session recovery and archiving
5. Future extensibility: Framework supports adding tools, retrieval, or multi-agent patterns if needed

Alternatives considered: Direct API calls would reduce dependencies but require reimplementing conversation history management, prompt templating, and cross-provider compatibility layers.

4.4.6. LangChain Limitations in This Context

- No chains used: Despite name ConversationChain, this is a direct LLM wrapper (no LangChain Expression Language chains)
- No tools/agents: Simple request-response pattern (could extend for future tool-use capabilities)
- Custom JSON parsing: LangChain's built-in output parsers not used; custom extraction handles malformed responses more robustly

4.4.7. Speech Processing

Speech-to-Text (STT):

- Provider: Deepgram Nova-2 (`deepgram-sdk` 4.8.1)
- Model: `nova-2` with US English (`en-US`)
- Smart formatting enabled
- Interim results for real-time partial transcription
- Voice Activity Detection (VAD) events
- Adaptive endpointing: 200ms (conversational stages) / 500ms (log-reading task)
- Utterance end timeout: 1000ms (conversational) / 2000ms (log-reading)
- Audio processing: RTSP stream from Misty → FFmpeg MP3 encoding → Deepgram WebSocket

Text-to-Speech (TTS) - Three options:

1. **Misty Onboard TTS** (this is the one we used): Native robot voice via onboard TTS
2. **OpenAI TTS**:
   - Model: `tts-1` (low-latency variant)
   - Voice: `sage`
   - Format: MP3, served via HTTP (port 8000)
   - Ultimately chose not to use because we wanted a more robotic, non-human voice
   - Didn't want the human voice influencing trust on its own (future research could look at trust in relation to type of voice)
3. **Deepgram Aura**:
   - Model: `aura-stella-en` (conversational female voice)
   - Format: MP3, served via HTTP
   - Ultimately chose not to use because we wanted a more robotic, non-human voice

4.4.8. Emotion Detection

Model: DistilRoBERTa-base fine-tuned on emotion classification

- HuggingFace identifier: `j-hartmann/emotion-english-distilroberta-base`
- Framework: `transformers` (4.57.1) pipeline
- Hardware: CUDA GPU acceleration (automatic fallback to CPU)
- Output classes: joy, anger, sadness, fear, disgust, surprise, neutral
- Mapped to interaction states: positively engaged, irritated, disappointed, anxious, frustrated, curious, neutral

4.4.9. Multimodal Robot behaviour

Expression System: 25 custom action scripts combining:

- LLM was prompted to choose an appropriate expression from a predefined set based on context.
- Facial displays (image eye-expression files on screen)
- LED color patterns (solid, breathe, blink)
- Arm movements (bilateral position control)
- Head movements (pitch, yaw, roll control)

Nonverbal Backchannel behaviours (RESPONSIVE mode only):

- Real-time listening cues triggered by partial transcripts (disfluencies, hesitation markers)
- Emotion-matched expressions (e.g., "concern" for hesitation, "excited" for breakthroughs)

LED State Indicators:

- Blue (0, 199, 252): Actively listening (microphone open)
- Purple (100, 70, 160): Processing/speaking (microphone closed)

4.5. Data Collection

Database: DuckDB relational database (`experiment_data.duckdb`)

Logged Data:

1. **Sessions table**: participant ID (auto-incremented P01, P02…), condition assignment, timestamps, duration

2. **Dialogue turns table**: turn-by-turn user input, LLM response, expression, response latency (ms), behavioural flags

3. **Task responses table**: submitted answers with timestamps and time-on-task

4. **Events table**: stage transitions, silence check-ins, timer expirations, detected emotions

4.6. Interaction Dynamics

4.6.1. Silence Handling

Silence detection: 25-second threshold triggers check-in prompt

- RESPONSIVE: "Still working on it? No rush - I'm here if you need help!"
- CONTROL: "I am ready when you have a question."

4.6.2. Emotion-Responsive behaviours (RESPONSIVE condition only)

Frustration tracking:

- Consecutive detection of frustrated/anxious/irritated/disappointed states
- Threshold: 2 consecutive frustrated turns triggers proactive support
- RESPONSIVE adaptation: "This part can be tough. Want me to walk you through it?"

Positive emotion matching:

- Celebratory language for curious/engaged states
- Momentum maintenance: "Yes! Great observation!"

Run Mode: Set programmatically in `mistyGPT_emotion.py` line 126:

```
RUN_MODE = "RESPONSIVE"  # or "CONTROL"
```

4.7. Prompt Engineering

Modular prompt system (PromptLoader class):

- core_system.md: Task framing, role description, output format schema
- role_responsive.md / role_control.md: Condition-specific personality instructions
- stage1_greeting.md through stage5_wrap_up.md: Stage-specific task instructions.

Context injection: Real-time contextual variables passed to LLM:

- Current stage
- Detected emotion (if enabled)
- Task submission status
- Timer expiration notifications
- Silence check-in flags

4.8. Inter-process Communication

Flask REST API endpoints:

- GET /stage_current: Synchronize stage state with facilitator GUI
- GET /task_submission_status: Detect participant task submissions
- GET /timer_expired_status: Detect timer expirations
- POST /stage: Update stage (facilitator override)
- POST /reset_timer: Clear timer expiration flags

References

Lin, T.H., Ng, S., Sebo, S., 2022. 2022 31st ieee international conference on robot and human interactive communication (ro-man), pp. 37–44. URL: https://ieeexplore.ieee.org/document/9900828, doi:10.1109/RO-MAN53752.2022.9900828. iSSN: 1944-9437.