Trust in Autonomous Human–Robot Interaction

An In-Person Pilot Study

M.C. Lau
Laurentian University
mclau@laurentian.ca

Shauna Heron
Laurentian University
sheron@laurentian.ca

2025-12-14

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.

Human–Robot Collaboration (HRC) has emerged as a critical area in the engineering and social sciences domain. The current paper ventures into this growing domain with a focus on an environment where collaboration pivots on problem-solving through shared knowledge and dialogue. Specifically, we explore collaboration performance and trust perception after interaction with two versions of a social robot capable of carrying out autonomous actions and decision-making under the guidance of a computer program. One version designed to be responsive and proactive to participants affect and queries and the other designed to offer help only when asked.

Trust is a central construct in human–robot interaction (HRI), shaping how people collaborate with, rely on, and accept robotic systems across social, assistive, and task-oriented domains [1]. In any kind of collaborative setting, including HRI, trust has been identified as a significant factor that can work to support or hinder cooperation, particularly in contexts characterized by incomplete or uncertain information [2]. Trust influences not only subjective evaluations of robots but also objective outcomes like task performance, compliance, and engagement [3]. As a result, a growing body of work has focused on evaluating trust following human–robot interactions, including the development of several standardized instruments designed to capture users' perceptions of robot reliability, predictability, and intent in various industrial, medical and social settings [2], [3], [4], [5].

Despite this growing literature, much of what is currently known about trust in HRI has been derived from interactions conducted under simulated conditions. In many studies, robot behavior is scripted, simulated by a computer program, or mediated through human control of a robot using Wizard-of-Oz (WoZ) paradigms [6], [7]. While such approaches are valuable for early-

stage design and hypothesis generation, these approaches critically alter interaction dynamics by masking real-world sensing failures, response latency, and behavioral inconsistencies that are characteristic of autonomous robotic systems. This gap is especially notable given that autonomy-related challenges—such as speech recognition errors, model hallucinations, delayed responses, and misinterpretations of user intent—are likely to play a critical role in shaping trust during real deployments. From an HRI perspective, understanding trust in the presence of real-world imperfections may be more informative than evaluations conducted under idealized assumptions. Nevertheless, few studies have directly examined trust outcomes following fully autonomous, in-person human–robot interaction.

To address this gap, the current study leveraged a between-subjects design to evaluate trust following an in-person interaction with a robot operating autonomously within predefined behavioral constraints. Participants collaborated with the robot in solving an immersive puzzle game where the robot served as a diegetic "game guide" and collaborative partner. In the game, the participant solved the mystery by interacting with the game guide to obtain hints, moral support and advice on how to solve puzzles; the robot managing speech-based interaction, task progression, and affect-responsive behavior, all without human intervention.

To achieve this we developed an autonomous spoken-language interaction system integrated with a speech recognition (ASR) pipeline, including affect detection with the Misty-II robot platform to allow the robot to engage in natural conversations with users. The system is capable of recognizing speech, managing dialogue, remembering what was said previously, and generating spoken responses and facial expressions and head and arm movement of the robot during dialogue.

By examining post-interaction trust using established trust measures alongside behavioral and task-level outcomes, this study aims to contribute empirical evidence on how trust might be shaped in fully autonomous HRI scenarios. Rather than seeking to demonstrate optimal performance under ideal conditions, the focus is on understanding trust as it is impacted during realistic human–robot interaction, where uncertainty, interactional breakdowns, and adaptive behavior are unavoidable. As such, this work provides insight into the practical implications of affect-responsive autonomy for trust in human–robot collaboration.

Hypothesese

Methods

Sample and recruitment

Participants (n = 29) were recruited from the Laurentian University community through word of mouth and via the SONA recruitment system. Eligible participants were adults (18+), fluent in english with normal to corrected hearing and vision and no 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 was purchased with grant funding from the IAMGOLD President Innovation Fund.

Sample characteristics are summarized in Table 1.

Experimental design

Participants interacted with the Misty-II robot in a shared physical workspace that included both the robot and a computer-based task interface. The interface was visible to participants and used to present brief task instructions, collect responses, and advance between task stages. Importantly, the robot autonomously monitored task progression and participant input through the interface, allowing it to adapt its dialogue and responses without human intervention.The interface served as a communication channel between the participant and the autonomous system rather than as a mechanism for remotely controlling robot behavior (See Figure 1).

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.

Participants collaborated with a Misty-II robot in solving an immersive puzzle game where the robot served as a diegetic "game guide" and collaborative partner. In the game, the participant solves the mystery by interacting with the robot guide for hints, emotional support and advice on how to solve the puzzles. The game was composed of two sequential tasks designed to elicit interaction with the robot under differing knowledge and dependency conditions [7] . The robot autonomously monitored task progression through the interface and adapted its dialogue accordingly without real-time human intervention.

In the control condition the robot followed a neutral interaction policy, while in the experimental condition the robot was prompted to adapt its behavior based on detected user affect, dialogue and

demands of the task itself. Importantly, both conditions utilized the same robot and were subject to the same sensory and interaction limitations inherent to autonomous operation, including speech recognition variability and response timing constraints. The only difference was the interaction policy between conditions.

Between-subjects factor: Robot Interaction Policy

1. **RESPONSIVE** (experimental): Warm, emotionally engaged, proactive behavior with emotion-responsive adaptation
2. **CONTROL** (baseline): Neutral, reactive, information-only responses

Task Structure

Five sequential stages:

1. **Greeting** (stage1): Participant introduction and rapport building
2. **Mission Brief** (stage2): Task explanation and scenario framing
3. **Task 1** (stage3): Who dunnit task
   - Identify suspect from 4×6 grid (24 options) by asking robot Yes/No questions based on ground-truth
   - Ground truth features: red hair, glasses, no hat, long hair, pink hoodie
4. **Task 2** (stage4): Log analysis task
   - determine missing robot location by collaborating with Misty-II to decipher system, wifi and sensor logs
5. **Wrap-up** (stage5): Debriefing and conclusion

Time constraints: ~15 minutes total session duration

Task 1: Robot-dependent collaborative reasoning

The first task required participants to identify a suspect from a 6x4 grid of 'suspects' by asking a series of yes/no questions about their features. A grid of potential suspects was displayed on the interface, and participants formulated questions verbally to narrow down the correct individual (e.g., 'was the suspect wearing a hat?'). In this task, the robot possessed the ground-truth information necessary to determine whether each question was true or false, making successful task completion dependent on interaction with the robot.

This task was designed to establish an initial forced collaborative dynamic in which the robot served as an essential informational partner. Participants were required to engage verbally with the robot and coordinate question strategies to reach a solution within the allotted time (5 minutes). The structured nature of the task ensured that the robot's role was clear and that collaboration was unavoidable.

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

The second task involved a more complex problem-solving scenario in which participants had access to multiple technical logs presented via a simulated terminal interface to determine the location of the missing 'Atlas' robot. Unlike the first task, the robot did not possess ground-truth knowledge about the whereabouts of the robot. The robot's assistance in this task was limited to

general problem-solving support derived from language model's prior training, such as explaining how to interpret log information, suggesting reasoning strategies, or helping participants reflect on inconsistencies across logs. The robot was explicitly constrained such that it was informed only that participants could view several logs, without access to the content of those logs or the correct answers to the task-related questions. The robot could ask the participant questions about what they found in the logs and the human could do the same.

Importantly, participants could complete the second task independently or choose to solicit assistance from the robot. As a result, the robot functioned as a collaborative reasoning partner rather than an authoritative source. Participants retained full control over decision-making and were free to accept, reject, or ignore the robot's suggestions. This design allowed collaboration to emerge voluntarily, rather than being enforced by task structure [7].

Wrap-up and debrief

Once all answers were submitted, the correct answers were shown to participants, letting them know how they did. At the wrap-up stage the robot and the participant had the chance to debrief on whether they were right or not, and then the task came to an end with the robot thanking them and prompting them to report to the principal investigator.

In-person procedure

Participants (n = 29) completed a pre-interaction questionnaire on Qualtrics where consent, demographics information, and a baseline measure of Negative Attitudes Towards Robot Scale and Need for Cognition (thinking style) were administered. Because of potential variability around timing of the pre-interaction tests and the in-person sessions, we elected not to use a formal pre-post test. Instead we took a general measure of attitudes towards robots as well as general thinking style to establish a baseline for later group comparison.

At the in-person session, once the pre-interaction survey was complete, participants were seated in front of Misty and instructed to start the session by clicking the Start button on the dash. They were also instructed on basic communication tips with the robot: i.e., to wait until the blue light on the side of the robot's head was on before speaking. Finally, once the participant was ready to start, the researcher left the room and closed the door, leaving the robot and participant to complete the tasks together. Once complete the participant would exit the room and then complete a post-interaction survey containing the Trust Perception-HRI scale and the Trust in Industrial Human-robot Collaboration scale followed by a written debriefing and verbal debriefing with the primary researcher. Participants were informed they could leave the room and stop the session at any time, no questions asked. Once complete, participants were presented with a $15.00 gift card as compensation for their time. All participants completed the tasks and remained for their full session which took an average of 30 minutes to complete.

Results

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).

Post-Interaction Trust Differences

Descriptive comparisons of participant-level post-test scores indicated an approximately 10 point difference in post-test Trust Perception Scale-HRI scores ($M \approx 73$ vs $M \approx 63$) and a 14 point difference in the Trust in Industrial Human-robot Collaboration scale ($M \approx 47$ vs $M \approx 61$) between conditions, although these differences did not reach conventional significance ($p<.05$) under a two-sample t-test ($p = 0.09$ and $p = 0.14$ respectively), reflecting our limited power ($n = 29$), Bayesian hierarchical models indicated higher post-interaction trust scores in the responsive robot condition across both trust-related scales (posterior mean differences ≈ 7–8 points on a 0–100 scale).

Although 95% credible intervals overlapped zero, the posterior probability that the responsive condition increased trust was 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 70%, suggesting a reasonable likelihood of a practically meaningful effect size. In a larger sample, these effects would likely reach conventional levels of statistical significance.

Trust subscale patterns

Examination of trust subscales suggested that group differences were most pronounced for affective components of trust (e.g., trust feelings), whereas differences in perceived reliability were smaller. This pattern aligns with correlations showing that baseline negative attitudes toward robots were more strongly associated with affective trust than with reliability judgments.

Interaction dynamics and task performance

Characteristic	N	CONTROL N = 13[1]	RESPONSIVE N = 16[1]	p-value[2]
post_trust	29	47 (26)	61 (26)	0.094
post_trust_reliability	29	46 (24)	62 (20)	0.11
post_trust_perception	29	49 (26)	57 (25)	0.32

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

[2] Wilcoxon rank sum test; Wilcoxon rank sum exact test; Pearson's Chi-squared test; Fisher's exact test

Characteristic	N	CONTROL N = 13[1]	RESPONSIVE N = 16[1]	p-value[2]
post_trust_feelings	29	59 (33)	72 (29)	0.26
Post-Task Trust Perception	29	63 (17)	73 (19)	0.16
Suspect ID Accuracy	29	4 / 13 (31%)	10 / 16 (63%)	0.089
Status Accuracy	29	11 / 13 (85%)	10 / 16 (63%)	0.24
building_correct	29	10 / 13 (77%)	12 / 16 (75%)	>0.99
floor_correct	29	10 / 13 (77%)	13 / 16 (81%)	>0.99
zone_correct	29	6 / 13 (46%)	4 / 16 (25%)	0.27
Total Task Accuracy	29	3.15 (0.99)	3.06 (1.34)	0.98
Overall Task Accuracy	29	0.63 (0.20)	0.61 (0.27)	0.98
Dialogue Turns	29	32 (10)	36 (11)	0.95
Avg Task Duration (mins)	29	12.84 (4.02)	16.81 (6.38)	**0.050**
Avg Response Time (ms)	29	15.1 (4.1)	17.2 (2.4)	**0.006**
Silent Periods	29	5.15 (2.27)	5.31 (2.82)	0.88
Engaged Responses	29	1.92 (2.25)	3.50 (1.83)	**0.020**
Frustrated Responses	29	0.54 (0.66)	0.88 (1.15)	0.56
n_neg	29	1.00 (0.91)	0.94 (1.18)	0.63

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

[2] Wilcoxon rank sum test; Wilcoxon rank sum exact test; Pearson's Chi-squared test; Fisher's exact test

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.

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 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 behavioral 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.

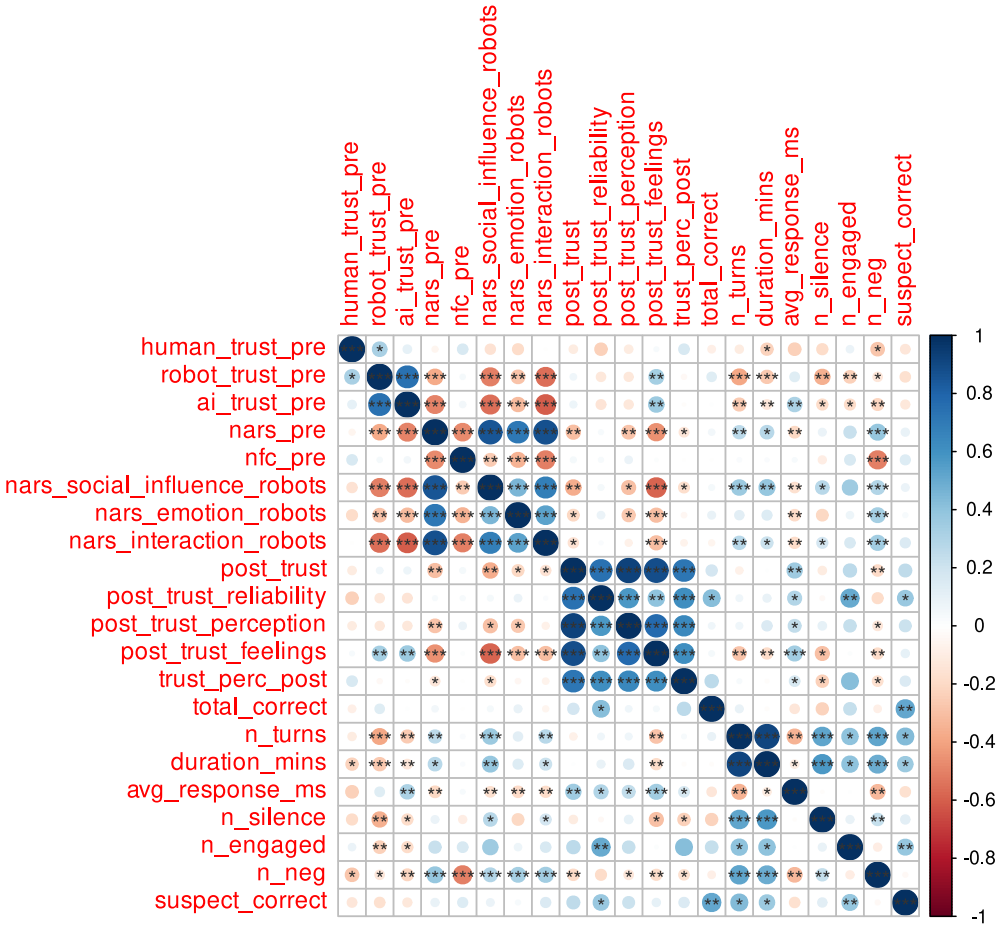

Figure 2

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

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 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 behaviors altered the affective tone of the interaction itself.

Appendix

Table 1: Participant Demographics and Baseline Characteristics by Group

Characteristic	N	CONTROL N = 13[1]	RESPONSIVE N = 16[1]	p-value[2]
Gender	27			0.84
Woman		6 / 13 (46%)	7 / 14 (50%)	
Man		7 / 13 (54%)	7 / 14 (50%)	
Age Group	27			0.35
18-24		5 / 13 (38%)	7 / 14 (50%)	
25-34		4 / 13 (31%)	2 / 14 (14%)	
34-44		1 / 13 (7.7%)	4 / 14 (29%)	
45+		3 / 13 (23%)	1 / 14 (7.1%)	
Program	25			>0.99
Psychology		1 / 13 (7.7%)	1 / 12 (8.3%)	
Engineering		2 / 13 (15%)	1 / 12 (8.3%)	
Computer Science		7 / 13 (54%)	6 / 12 (50%)	
Earth Sciences		0 / 13 (0%)	1 / 12 (8.3%)	
Other		3 / 13 (23%)	3 / 12 (25%)	
Experience w/Robots	29	7 / 13 (54%)	4 / 16 (25%)	0.14
Native English Speaker	29			0.53
Native English		5 / 13 (38%)	8 / 16 (50%)	
Non-Native English		8 / 13 (62%)	8 / 16 (50%)	
NARS Overall	29	38 (8)	38 (7)	0.79
Need for Cognition	29	3.62 (0.78)	3.74 (0.74)	0.55

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

[2] Pearson's Chi-squared test; Fisher's exact test; Wilcoxon rank sum test

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;

Technical Specifications

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.

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

Software Architecture

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

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.

LangChain Framework Integration

Core LangChain Components

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

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

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

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 behavior)
- 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

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...")

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.

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

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

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

Multimodal Robot Behavior

Expression System: 25 custom action scripts combining:

- 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 Behaviors (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)

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), behavioral 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

Additional logs:

- Prompt debugging logs (full LLM prompts with history)

- Audio recordings (MP3, timestamped by utterance)

- Real-time console logs with structured event markers

Interaction Dynamics

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."

Emotion-Responsive Behaviors (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!"

System Requirements

Hardware:

- CUDA-capable GPU (optional, for emotion detection acceleration)
- Network: Local HTTP server (port 8000 for audio serving, port 5000 for Flask GUI)
- Misty II robot on same local network

Software:

- Python 3.10
- PyTorch 2.9.0 with CUDA 12.8 support
- FFmpeg system installation

Configuration

Environment Variables (`.env` file):

- `GPT_API_KEY`: OpenAI API key (required if using GPT or OpenAI TTS)
- `DEEPGRAM_API_KEY`: Deepgram API key (required)
- `GEMINI_API_KEY` / `GOOGLE_API_KEY`: Google Gemini API key (required if using Gemini)
- `LLM_PROVIDER`: "GEMINI" or "OPENAI" (default: GEMINI)
- `TTS_PROVIDER`: "misty", "openai", or "deepgram" (default: misty)
- `ENABLE_EMOTION_DETECTION`: true/false (default: true)

- `DEBUG_PROMPTS`: 0/1 - Enable full prompt logging
- `LLM_TEMPERATURE`: 0.0-1.0 (default: 0.7)
- `WRAP_TIMEOUT_SECONDS`: Auto-exit timeout (default: 60)

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

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

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

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

Reliability Features

- Graceful shutdown handlers (SIGTERM, SIGINT)
- Automatic session logging even on interruption
- Fallback behaviors: CPU-only emotion detection, Misty TTS if cloud services fail
- Deepgram speech_final detection with UtteranceEnd backup handler

Bibliography

[1] E. Loizaga, L. Bastida, S. Sillaurren, A. Moya, and N. Toledo, "Modelling and Measuring Trust in Human–Robot Collaboration," *Applied Sciences*, vol. 14, no. 5, p. 1919, Jan. 2024, doi: 10.3390/app14051919.

[2] K. E. Schaefer, "Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI," R. Mittu, D. Sofge, A. Wagner, and W. Lawless, Eds., Boston, MA:

Springer US, 2016, pp. 191–218. [Online]. Available: https://doi.org/10.1007/978-1-4899-7668-0_10

[3] G. Charalambous, S. Fletcher, and P. Webb, "The Development of a Scale to Evaluate Trust in Industrial Human-robot Collaboration," *International Journal of Social Robotics*, vol. 8, no. 2, pp. 193–209, Apr. 2016, doi: 10.1007/s12369-015-0333-8.

[4] I. Cucciniello, S. Sangiovanni, G. Maggi, and S. Rossi, "Mind Perception in HRI: Exploring Users' Attribution of Mental and Emotional States to Robots with Different Behavioural Styles," *International Journal of Social Robotics*, vol. 15, no. 5, pp. 867–877, 2023, doi: 10.1007/s12369-023-00989-z.

[5] S. Diefenbach, M. Herzog, D. Ullrich, and L. Christoforakos, "Social Robot Personality: A Review and Research Agenda," Springer VS, Wiesbaden, 2023, pp. 217–246. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-658-37641-3_9

[6] R. Maure and B. Bruno, "Autonomy in socially assistive robotics: a systematic review," *Frontiers in Robotics and AI*, vol. 12, p. 1586473, doi: 10.3389/frobt.2025.1586473.

[7] T.-H. Lin, S. Ng, and S. Sebo, "2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)," Aug. 2022, pp. 37–44. doi: 10.1109/RO-MAN53752.2022.9900828.