



A Hybrid LLM/Model-Based Architecture for Flexible and Adaptive Robot Coaching

Luigi Gargioni^{1,3} · Rachid Alami² · Daniela Fogli¹

Received: 24 September 2025 / Revised: 21 May 2026 / Accepted: 13 July 2026
© The Author(s) 2026

Abstract

Human-Robot Interaction (HRI) presents significant challenges in accurately assessing situations, adapting robotic behavior to human intentions, ensuring explainability, pertinence, and acceptability, and effectively managing uncertainty. Traditional model-based approaches provide reliability but struggle with human unpredictability, often approximating human behavior through specific models that fail to encompass all possible scenarios. Conversely, statistical approaches, such as Large Language Models (LLMs), offer greater adaptability but lack deterministic guarantees. This paper introduces a hybrid architecture that integrates structured methodologies with the flexibility of LLMs to enhance robotic coaching in dynamic environments. The proposed architecture leverages deterministic modules for critical constraints and safety guarantees while employing LLMs for contextual understanding, natural language interaction, and adaptation to unpredictable human behaviors. Through a healthcare robotic coach scenario implementation and the conduction of a user study, we aim to demonstrate how this balanced approach enables effective monitoring of task execution, dynamic adaptation to human states, and seamless verbal interaction while maintaining system reliability. By bridging deterministic and statistics-based techniques, the proposed architecture aims to advance HRI toward safer, more transparent, flexible, and adaptive interactions.

Keywords Human-robot collaboration · Robot coaching · Large language model · Model-based · Hybrid architecture

1 Introduction

Human-Robot Interaction (HRI) is a multidisciplinary research domain focused on studying and designing interactions between humans and robots. This field integrates principles from artificial intelligence, robotics, cognitive

science, psychology, ethics, and human factors engineering to develop systems that facilitate natural, effective, and intuitive interaction between humans and autonomous machines. As robots become increasingly embedded in various aspects of daily life, ranging from industrial automation [1, 2] to personal assistance and healthcare [3, 4], ensuring seamless human-robot interaction is of paramount importance. However, HRI presents challenges that extend beyond those encountered in conventional automation. It necessitates close cooperation between humans and robots to achieve shared objectives, requiring sophisticated reasoning, planning, and adaptive mechanisms to enable fluid interaction and effective task execution. A fundamental challenge in HRI lies in equipping robots with the capability to infer and estimate human beliefs and intentions and adjust their behavior accordingly. Unlike traditional automation, where robot actions follow predefined rules, effective HRI requires robots to dynamically interpret and respond to human actions, preferences, and situational changes. This entails not only processing explicit commands but also recognizing implicit cues such as gestures, emotions, and

✉ Luigi Gargioni
luigi.gargioni@unimib.it

Rachid Alami
rachid.alami@laas.fr

Daniela Fogli
daniela.fogli@unibs.it

¹ Department of Information Engineering, University of Brescia, Via Branze 38, 25123 Brescia, Italy

² Robotics and InteractionS (RIS), LAAS-CNRS, 7 Av. du Colonel Roche, Toulouse 31400, France

³ Present address: Department of Psychology, University of Milano-Bicocca, Viale dell'Innovazione 10, Building U9, 20126 Milan, Italy

contextual information. Another critical aspect is the ability of robots to plan and coordinate their actions in a manner that ensures transparency, explainability, and acceptability. Robots should assist human collaborators in anticipating and comprehending robotic behavior, thereby minimizing unpredictability that could lead to confusion or inefficiency. Achieving this goal necessitates generating actions that align with human expectations while adhering to safety considerations, ethical constraints, and usability principles.

Traditional model-based architectures provide several important strengths that make them fundamental in robotics. Their reliance on explicit models and formal reasoning ensures predictability, verifiability, and safety guarantees, which are crucial in safety-critical domains such as industrial automation and healthcare. Furthermore, model-based methods facilitate transparent decision-making and explainability, since the reasoning process and control rules can be explicitly traced and validated. These characteristics make deterministic architectures particularly valuable when reliability and accountability are paramount. However, despite these advantages, their dependence on predefined models and static rules limits adaptability to dynamic human behaviors and context-aware interactions.

The integration of Large Language Models (LLMs) presents a promising avenue to tackle these challenges. These models significantly broaden the spectrum of scenarios that robots can manage, enabling more flexible and larger spectrum approaches to situation assessment and behavior synthesis. By leveraging common sense reasoning and contextual understanding, LLMs can assist robots in interpreting human inputs and environmental cues, inferring implicit intentions, and generating adaptive responses that align more closely with human expectations. Nevertheless, an approach based only on statistical models remains inadequate for ensuring reliability and safety in HRI. It is essential to develop an architecture in which the generated actions are validated and the critical aspects (such as safety constraints and ethical considerations) are explicitly defined and enforced.

Given the respective strengths and limitations of traditional architectures and model-based methodologies, as well as the emerging capabilities and challenges of LLMs, this paper proposes a hybrid architecture that integrates the structured reliability of deterministic models with the adaptability, common sense reasoning, and contextual awareness offered by LLMs. In this context, the term *hybrid* denotes the combination of deterministic techniques, providing formal guarantees and rule-based precision, with non-deterministic, statistics-based models that enable learning, adaptability, and robustness in uncertain environments.

In this work, we investigate the extent to which a hybrid LLM/model-based architecture can enhance task execution,

adaptability, and user interaction in HRI. We hypothesize that integrating LLMs supports adaptability to human unpredictability and decision-making efficiency, and fosters natural and intuitive interactions while maintaining safety and reliability through deterministic components. This leads us to our central research question (RQ): *How can a hybrid LLM/model-based architecture support task execution, adaptability, and user interaction in HRI while ensuring safety, reliability, pertinence, and transparency?*

By integrating these complementary paradigms, we aim to develop a flexible and human-aware HRI system capable of operating effectively in diverse real-world settings while maintaining verifiable safety and transparency.

A preliminary and partial version of this architecture was presented as a workshop paper in [5]. The aim was to collect early feedback on the proposed idea; the present paper substantially extends it with a complete architecture, implementation, and empirical evaluation.

2 Related Work

Several studies in the literature have proposed control architectures to develop adaptive robots able to effectively interact with humans [6, 7]. These architectures are designed to integrate decisional and functional components within a unified framework that efficiently manages information flow. Decisional components typically encompass higher-level processes such as situation assessment and planning, which are essential for enabling robots to make informed decisions based on human presence and environmental context [8, 9]. Conversely, functional components, including perception and action, allow robots to engage with the physical world and dynamically respond to their surroundings in real-time [10, 11].

The primary challenge lies in identifying, structuring and articulating these diverse components into a cohesive architecture capable of addressing the system's complex requirements. This involves identifying the processes required to accomplish each task, defining their temporal interdependencies to ensure coherent execution, and assigning to each component specialized decisional or functional mechanisms to effectively address the specific challenges of HRI that may arise. A key limitation of these architectures is their reliance on predefined models of human beliefs, intentions, and preferences. Given the inherent unpredictability of human behavior, accurately modeling these aspects remains a significant challenge. This constraint limits the system's capacity to fully anticipate and adapt to human situations and actions, introducing a layer of uncertainty that remains difficult to mitigate. The effective integration of these components ensures that the robot can handle both the strategic

and tactical aspects of its actions, adapting to dynamic environments and interacting with humans in meaningful ways.

To address these challenges, cognitive architectures incorporating belief management and Theory of Mind (ToM) offer a more flexible approach, enabling robots to model, interpret, and act appropriately to the mental states of human agents [12, 13]. In HRI, different actors possess distinct background knowledge, varying levels of awareness regarding the task and environment, and potentially conflicting perspectives. To navigate this complexity, a robot must continuously maintain, update, and reason about diverse belief systems, ensuring that the most current and relevant information informs its actions. Belief management frameworks play a fundamental role in this process, enabling the robot to infer, track, and reconcile different viewpoints before making decisions or executing actions [14–16].

Beyond control architectures, various frameworks have been developed to support the generation of action plans for collaborative tasks between humans and robots [17]. These planning systems enable robots to formulate effective action strategies while operating alongside human partners. A primary challenge in these systems lies in ensuring that the generated plans account for both the robot's capabilities and the human's behaviors and expectations. Most planning systems, including those incorporating adaptation or learning mechanisms, are inherently model-based [18]. They depend on predefined models representing the human, the task, and the interaction dynamics between the human and the machine. These models are essential for predicting human actions, inferring intentions, and anticipating task evolution. However, given the intrinsic unpredictability of human behavior, uncertainty remains a critical factor in these systems, necessitating advanced methods to address it.

Markov Decision Processes (MDPs) [19] and Partially Observable Markov Decision Processes (POMDPs) [20] are commonly employed to model uncertainty in observation and to estimate action outcomes. MDPs are particularly useful when action results are uncertain, requiring decision-making that balances immediate and future rewards. POMDPs extend MDPs by considering scenarios in which the robot has incomplete information about the environment or the human's state or intention (the latter is an essential factor in human-robot interaction). MDPs and POMDPs models enhance decision-making by mitigating the uncertainty inherent in human behavior and environmental variability.

Epistemic planning explicitly models the divergence in beliefs between the robot and the human [21]. By incorporating each agent's knowledge of the other's beliefs and intentions, epistemic planning facilitates coordination, fostering seamless interaction between humans and machines. Additional approaches include non-deterministic planning

frameworks, where human decisions and actions are treated as contingencies [22]. Ethical planning is another noteworthy paradigm, ensuring that the generated plans adhere to ethical principles and constraints [23].

Another significant challenge in these planning schemes is modeling human cognitive processes. The ACT-R/E (Adaptive Character of Thought-Rational/Embodied) model [24] has been used to simulate human cognitive processes, providing a framework for predicting and understanding how humans approach tasks and make decisions. Incorporating such cognitive models into the planning process enhances the robot's ability to anticipate human behavior, allowing for more effective cooperation.

To ensure that the robot's actions are both acceptable and pertinent, several criteria have been proposed, often framed in terms of costs and utility, to assess the efficiency and relevance of different action plans. Frameworks such as HATP (Human Aware Task Planning) [25] have been used to define these criteria, guiding the robot's decision-making process by considering both the practical costs of actions (e.g., time, resources) and their expected benefits (e.g., effectiveness, satisfaction). This approach allows robots to plan actions that align with human preferences and expectations, optimizing collaborative task achievement. By integrating these advanced methodologies, robots can better adapt to the uncertainties of human behavior, improve task execution in dynamic environments, and ensure their actions align with the collaborative goals of human-robot interaction.

Despite significant advancements, existing approaches to HRI continue to struggle with the fundamental challenge of mind reading and the inherent difficulty of accurately representing human beliefs and decision-making processes. This limitation makes it challenging for robots to interpret complex or nuanced human intentions. The effectiveness of these methodologies relies on the accuracy of models representing the environment, the task, and the strategies for task execution. However, constructing a comprehensive and robust model of the human remains a major obstacle due to the unpredictability of human behavior. The vast range of possible scenarios raises concerns about the viability of relying exclusively on predefined models for human-robot interaction.

Recent contributions have highlighted the potential of LLMs in addressing the above challenges. In [26], LLMs are described as a "paradigm shift" for robotic systems, enabling unprecedented levels of natural language understanding and task execution. The integration of LLMs further enhances robots' ability to solve linguistic ambiguities and infer implicit user intentions. In [27], an adaptive framework in which an LLM acts as a bridge between the user and the robot is proposed, tailoring interaction strategies to user characteristics such as age; for example, with older adults,

the system simplifies dialogue and makes robot intentions more explicit, thereby improving usability and transparency. Recent work has further demonstrated the potential of LLMs specifically in health-oriented coaching scenarios: in [28], LLM-powered socially assistive robots are shown to effectively promote cognitive health in elder care through clinically-grounded verbal interaction, while in [29], an LLM-based physical activity coaching system demonstrates personalized support capabilities, though also highlighting the risks specific to LLM-based health coaching.

Despite these advantages, significant limitations remain for LLMs in HRI. In [30], it is reported that the statistical basis of LLMs, combined with the absence of formal guarantees, poses risks to safety in real-world deployments. In particular, LLM outputs may suffer from factual inaccuracies, and more critically, LLM-driven task planning often lacks temporal coherence, leading to unstable or unsafe robot trajectories and actions. In [31], it is discussed that while LLMs improve the naturalness of interaction, classical symbolic planners still outperform them in terms of reliability and efficiency of task planning.

These findings are further corroborated by a recent systematic review [32], which confirms that reliability, safety, and evaluation remain major open challenges in LLM-based HRI. The need to address these challenges has motivated the emergence of neuro-symbolic approaches that explicitly combine the flexibility of neural models with the formal guarantees of symbolic methods. In [33], a neuro-symbolic framework bridging LLMs and PDDL-based task planning is proposed for human-robot collaboration scenarios, while acknowledging that LLM-based planners should not yet fully replace traditional symbolic approaches. Similarly, in [34], it is argued that for humanoid assistive robots, critical safety and access-control constraints require a hybrid approach that pairs neural planning with symbolic enforcement at runtime. From a user perspective, in [35] it is shown empirically that LLM-powered robots excel in connection-building but underperform in logical communication, further motivating architectures that combine LLMs with structured deterministic components.

In summary, while it is evident that LLMs substantially advance natural dialogue and adaptability, they remain insufficient for ensuring robust, safe, and temporally consistent robot actions. Therefore, this paper proposes integrating LLMs with a model-based approach.

3 The Proposed Hybrid Architecture

HRI presents various challenges, ranging from deterministic task execution to highly dynamic and ambiguous human behaviors. In this context, the goal is defined and shared

with the human (i.e., Human-Robot Joint Action [36]), ensuring that both the robot and the human share a common understanding of the desired outcome and can contribute to reaching it. This shared goal guides the robot's actions and decisions, aligning its behavior with human expectations. Furthermore, a task refers to a specific activity that the robot and the human have to perform to achieve a specific goal, which can vary in complexity from simple, predefined operations to more adaptive and interactive behaviors requiring real-time situation assessment. The decision process is essential because, even if the task is well specified, there are different ways to achieve the same goal.

Nevertheless, HRI is predicated on certain assumptions that must be examined. The human and the robot are inherently different entities. The robot's purpose is to assist the human in effectively accomplishing a task (i.e., *robot coaching*). Furthermore, the human is regarded as an unpredictable entity, yet it is presumed that they will collaborate with the robot and will not deceive it with malicious intent. Both the robot and the human participate in the task, employing multimodal (i.e., speaking and acting) interaction to support the successful achievement of the objective.

Altogether, the challenges and assumptions discussed above underscore the importance of a structured approach, leading to the hybrid LLM/model-based architecture for HRI proposed in this paper and shown in Fig. 1.

The architecture is composed of interconnected modules that work collaboratively to ensure seamless interaction, reliable task execution, and robust adaptability to evolving scenarios.

Complementing these modules, a *Vector Database* serves as the central knowledge repository. It stores embeddings derived from unstructured data, including information about the user, the task, and the environment. Furthermore, it is continuously enriched with knowledge from prior interactions, such as user preferences inferred from past interactions. This repository supports task definition and monitoring progress by providing relevant contextual information.

In the following, it is shown which modules use a deterministic approach (i.e., green rectangles) and which use an LLM (i.e., yellow hexagons), and the reasons for this are explained in each case.

3.1 Knowledge Manager

This module leverages an embedding model to generate embeddings from natural language text, such as text files. Beyond extracting semantic representations of document content, it also appends relevant metadata, such as user identifiers, to each embedding. This ensures that the retrieved information remains user-specific and contextually relevant. The generated embeddings are then stored in

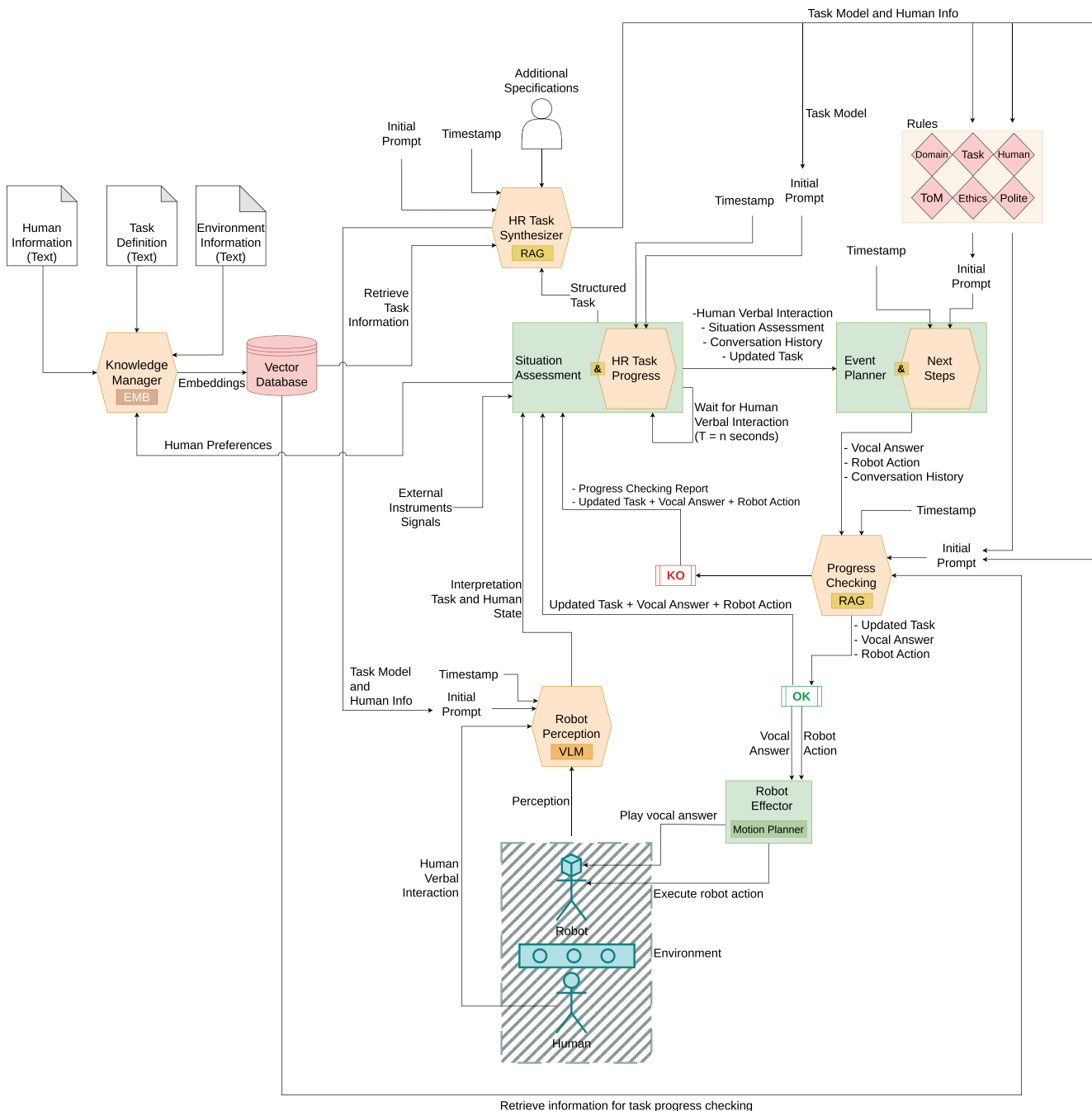


Fig. 1 The proposed hybrid LLM/model-based architecture

the *Vector Database* as a foundation knowledge for subsequent processing steps.

Additionally, when a daily task is completed, the *Knowledge Manager* is responsible for analyzing the interaction history by identifying key insights and extracting user preferences. These insights are encoded into embeddings and integrated into the *Vector Database*, ensuring continuous system adaptation based on newly acquired user-specific knowledge. To enhance the accuracy and relevance of the

stored information, an LLM is employed during the information extraction phase from the interaction history.

3.2 Human-Robot Task Synthesizer

This module leverages an LLM and Retrieval-Augmented Generation (RAG) to convert unstructured information from the *Vector Database* into a structured task representation. The generated data structure encodes activities, temporal

constraints, priorities, execution conditions, and contextual information that are subsequently monitored and orchestrated during task execution. This transformation is crucial for transitioning from natural language descriptions to a structured task plan formatted in JSON that can be programmatically processed. Additional specifications, including timestamps, are incorporated to generate precise task representations. Furthermore, it is possible to specify additional information in natural language for that specific iteration. The JSON format has been selected due to its compatibility with LLMs, which, when appropriately instructed through function calls (usually called also *tool calling* or *function calling*), can generate precise, structured JSON objects. The information retrieved from the *Vector Database* is also useful to other architecture modules, making them context-aware of the task and the user.

3.3 Situation Assessment & Human-Robot Task Progress

This module combines a rule-based approach (implemented through a *Situation Assessment* component) with an LLM (*Human-Robot Task Progress* component). *Situation Assessment* functions serve as the high-level control unit, responsible for parsing the request of a structured task, interpreting the task and human state, and managing the history of interactions, including recorded verbal responses and robot actions. Additionally, it evaluates progress checking reports, ensuring task completion, and, if discrepancies arise, it refines the output of other modules. Moreover, it incorporates predefined conditions that can trigger immediate deterministic reactions without passing through the module responsible for deciding the next steps (i.e., *Event Planner & Next Steps*). For example, if the user attempts to alter the task or if the user skips a mandatory activity, these behaviors activate an immediate response to address the situation (e.g., calling a human supervisor). The activation of such responses is managed by the *Human-Robot Task Progress* component, which monitors whether the conditions are occurring. This module can also process signals from external instruments, allowing the system to react deterministically to specific parameters (e.g., temperature of the room).

Through a rule-based approach, the *Situation Assessment* component orchestrates the workflow by managing the execution of other modules. It verifies task completion, updates the *Vector Database* with newly acquired knowledge, such as user preferences and behavioral patterns, and ensures alignment between the human-robot interaction and task objectives. Furthermore, it governs the robot's behavior, allowing it to engage the human proactively or reactively, responding to human requests and environmental cues. A control loop with a predefined waiting period is

implemented to handle scenarios where both the robot and the human remain passive.

The *Human-Robot Task Progress* component employs an LLM to monitor task execution and ensure that the process adheres to predefined objectives. The LLM is initialized with the task model and the invocation timestamp, providing temporal awareness. Its primary function is to update the task's status based on information received from the *Robot Perception* module. The architecture's flexibility is reinforced by the absence of a rigid algorithm governing task progression, thus preventing excessive constraints and enhancing adaptability to dynamic conditions.

In summary, the *Situation Assessment & Human-Robot Task Progress* module is responsible for updating the state of the world, the task itself, and the human's beliefs, preferences, behaviors, and goals.

3.4 Event Planner & Next Steps

This module consists of an event planner and an LLM that determines subsequent actions and interaction strategies. The *Event Planner* is a deterministic component, specifically an expert system, responsible for managing the sequence of activities within a task in a feasible order, considering scheduled times with tolerances, priorities, criticalities, and durations. It also accommodates unscheduled activities, placing them in available time slots. Furthermore, if two activities with the same priority can be carried out, the user can choose between them.

The expert system is based on a classical architecture that includes a knowledge base (storing facts and if-then rules), a working memory that represents the current state of the system, and an inference engine that applies the rules to the working memory using a forward-chaining methodology to derive new facts or trigger specific actions. The current implementation relies on a lightweight custom rule-processing mechanism rather than on a large-scale production rule engine based on optimization algorithms. Given the relatively limited number of activities and constraints considered in the present healthcare scenario, this approach was sufficient for real-time execution. This structure allows the system to reason in a deterministic and explainable manner, ensuring transparency and predictability in decision-making.

More specifically, the *Event Planner* operates as a rule-based temporal scheduler and reactive executor since it coordinates predefined or dynamically updated activities according to temporal constraints, priorities, contextual conditions, and interaction events. This design deliberately prioritizes transparency and deterministic executability over search-based deliberation, making the module's behavior fully traceable and predictable at runtime. In addition,

proactive behavior is supported through continuous monitoring of the task state and temporal conditions. The current time and contextual information are periodically reassessed by the *Situation Assessment* and *Event Planner* modules, enabling the system to autonomously trigger reminders, initiate interactions, suggest activities, or adapt schedules when specific conditions become valid. For example, medication reminders can be automatically activated when the corresponding temporal window is reached, without requiring explicit user input. While this rule-based approach is effective for routine and well-defined scenarios, the module could be replaced in the future by an automated planner for more complex tasks involving conditional dependencies or long-term objectives. This planner could break down high-level objectives into atomic activities, sequence them according to temporal and logical constraints, and adapt dynamically as the state of the system evolves.

The *Next Steps* LLM is tasked with determining the robot's next action and generating a verbal response for the user, ensuring coherence with the planned sequence of activities. This model is initialized with a comprehensive prompt, including task-related instructions, contextual rules, and user-specific information. Furthermore, the prompt explicitly encodes ethical constraints, including respect for user autonomy, avoidance of coercion, and appropriate escalation policies in safety-critical situations.

It is important to note that the *Human-Robot Task Progress* and *Next Steps* LLMs have been intentionally implemented as separate components. This design choice reflects their fundamentally different temporal scopes and reasoning objectives. The *Human-Robot Task Progress* LLM operates with a focus on the present state of the task: it continuously analyses incoming data and user interactions to maintain an accurate, up-to-date representation of what has already occurred and what is currently happening. By contrast, the *Next Steps* LLM is forward-looking: it reasons about future actions, deciding how the task should progress from the current state toward its completion. Combining these two temporal perspectives into a single model might blur the distinction between factual, observed states and hypothetical, planned states. This could lead to temporal inconsistencies—for example, the model might conflate actions that have already been completed with those still pending, or make predictions based on outdated or incorrect assumptions about the present. Such confusion would undermine the reliability of the reasoning process and could result in suboptimal or even unsafe task execution.

3.5 Progress Checking

This LLM-based module ensures the validity of the plan, the robot's actions, and the verbal responses generated by the

system. It employs RAG to retrieve complementary information or verify existing knowledge from the *Vector Database*, ensuring compliance with predefined task properties. This mechanism is crucial for increasing robustness against potential hallucinations or inaccuracies that the LLM might produce by grounding its outputs in verified, contextually relevant data. Additionally, it performs a double-check on the constraints introduced in the initial prompt of the *Next Steps* module, reinforcing adherence to the specified guidelines. This iterative validation loop enhances task reliability and minimizes errors that may arise during task execution and the planning of subsequent steps.

3.6 Robot Effector

This module executes actions and facilitates communication with the user through vocal responses and physical interactions. It ensures the system's decisions are applicable in the real world while maintaining contextual responsiveness. For verbal communication, the robot can use its built-in speakers to deliver voice responses to the user. When executing physical actions in the environment, the robot relies on an integrated Motion Planner that enables precise and context-aware movement. This dual capability allows the robot to effectively engage with the user and adapt its actions based on situational requirements.

3.7 Robot Perception

This module integrates a Visual Language Model (VLM) to process visual data, extract situational information from the environment, and provide continuous feedback to the *Situation Assessment & Human-Robot Task Progress* module. It interprets images to assess the task context, human state, and environmental conditions, ensuring informed decision-making. Domain-specific knowledge, including task details and human-related information, is crucial for accurate situational understanding. The VLM's capability to leverage general knowledge and apply common sense enables it to generate meaningful insights that facilitate task progression.

4 Interaction Workflow

The system initialization begins with the collection of unstructured information provided in natural language. This information is converted into embeddings and subsequently stored in the *Vector Database*. At the beginning of the interaction workflow, the *Human-Robot Task Synthesizer* extracts and processes the relevant information to create a structured task represented in JSON format.

Interacting with other modules, the *Situation Assessment & Human-Robot Task Progress* module combines the structured task with situational data, previous task states, and the history of interactions, generating an updated task plan. This plan is continuously adapted and refined based on the evolution of the situation and newly acquired data.

As the task progresses, the system monitors its execution through the *Event Planner & Next Steps* module, which coordinates and validates subsequent steps, ensuring that the outcomes meet predefined success criteria. If the *Progress Checking* module identifies discrepancies between the expected and actual results, the task is updated and returned to the *Situation Assessment & Human-Robot Task Progress* module, creating a feedback loop for iterative refinement.

Finally, the *Robot Effector* module, which enables the robot to interact physically and verbally, enables the system to dynamically engage with the environment while maintaining safety and user-centric functionality. Combined with the *Robot Perception* module, it plays a crucial role by continuously interpreting the task and human states, ensuring that the system remains adaptable and contextually aware.

The workflow ends when the task is successfully completed. At this stage, the *Knowledge Manager* analyses the interaction that took place, extracting relevant insights about the user and updating the *Vector Database*. The updated knowledge will support more adaptive and context-aware behavior in future interactions.

5 Example Scenario

To better illustrate the interaction workflow, let us consider a scenario in a healthcare setting where an assistive robot is tasked with helping a patient adhere to a prescribed therapy. While healthcare applications typically require high reliability standards, this approach offers valuable benefits by prioritizing patient comfort and adaptability alongside adherence to a prescribed therapy. Medical environments are often unpredictable, and patients may have varying daily rhythms and responses to therapy. However, while it is important to accommodate the patient's needs and pace, critical tasks must be performed with precision and consistency. This scenario thus represents a compelling use case where flexibility is applied up to a point, enabling adaptability and patient-centered care, while maintaining rigidity where it is essential to ensure safety and therapeutic effectiveness.

The scenario unfolds as follows:

1. The daily therapy of the patient is specified by a medical doctor in natural language, along with additional instructions on how to follow it, such as “*It is morning*

before breakfast, human and robot are present in the room. Assist and ensure that the patient takes their daily medication as specified in the prescription.”. Optionally, the doctor could add some additional specifications for the day. For example, “*Avoid sugar today.*”, “*Skip Paracetamol today.*”, or “*Do physical exercise at 10:00 today.*”.

2. These inputs, saved in a simple text file, are processed into structured information using embeddings stored in the *Vector Database*, where they are combined with existing information about the patient and the context, which is also stored in the *Vector Database*.
3. The *Human-Robot Task Synthesizer* module exploits relevant details (e.g., medication type, timing constraints, assumption notes, etc.) and formulates a structured task (i.e., JSON structure).
4. The robot, acting as both a supportive and authoritative assistant, monitors task execution through the *Situation Assessment & Human-Robot Task Progress* and *Event Planner & Next Steps* modules. It dynamically adapts to the human's preferences and emotional state, guiding them progressively toward task completion. If the patient correctly takes the medication, the system validates the action and updates the task status. If a discrepancy is detected, such as the patient refusing or missing a dose, the robot assesses their commitment level and responds accordingly. It may initiate a gentle reminder (e.g., “*Taking it now will help you stay on track with your treatment, and it will make you feel better. Please take this pill.*”) or escalate to a more authoritative intervention, such as alerting a caregiver (e.g., “*The patient refuses to take their medication at the scheduled time. Please check it with them.*”). This adaptive strategy, informed by past interactions, ensures effective assistance while respecting the patient's autonomy.
5. The *Robot Perception* module continuously interprets the task and human states, ensuring real-time adaptability. For instance, it reports whether the patient has taken the medicine (e.g., “*The patient took the red pill.*”) or whether the patient is sleeping and can not reply or do anything.
6. The *Robot Effector* module executes the corresponding physical actions, such as referring to the medication in an accessible manner (e.g., referring to a pill by shape if the patient is color-blind) or guiding the patient through the assumption process (e.g., “*This medicine should be taken with water just before lunch.*”).

The scenario shows the need for flexibility to effectively manage the inherent unpredictability of a real-world setting where conflicting needs may arise. On the one hand, the critical nature of certain tasks, such as medication

adherence, makes them mandatory. On the other hand, the patient's behavior, emotional state, and daily condition can vary significantly; for example, the patient might be unwilling or unable to take medication at the scheduled time due to fatigue, confusion, or discomfort. Without adaptive flexibility, a rigid system might fail to respond adequately, causing stress for the patient or risking non-compliance. By blending deterministic rules for critical constraints with adaptive, context-aware communication, the system can both enforce necessary precision and accommodate variations in patient behavior. This balance ensures that critical tasks are performed safely while respecting the patient's autonomy and real-life variability, illustrating the essential role of flexibility in countering instability and delivering effective healthcare assistance.

Preliminary testing was conducted using this scenario with Ollama, employing Llama3.3-70B as the primary LLM for the different modules, and Llama3.1-8B for the *Knowledge Manager* to create embeddings and *Progress Checking* modules, running on an NVIDIA RTX 6000 Ada Generation. In addition, OpenAI GPT-4.1 was also tested, yielding more human-like responses. The temperature for the primary LLMs was set to 0.5 to make the responses more human and less robotic. Chroma was chosen as a vector database for this implementation due to its open-source nature and seamless integration capabilities. This modular approach was developed using LangChain and LangGraph. The developed architecture has been made available on GitHub¹.

The architecture is implemented with two exceptions: the *Robot Perception* module does not utilize a VLM, and instead, any VLM outputs are manually written; additionally, the Motion Planner within the *Robot Effector* module is not implemented, so the robot's actions are limited to text-based responses (e.g., MOVE_MEDICATION_X_TO_PATIENT). The interface for communicating with the robot now relies solely on written commands, meaning that all interactions, including instructions and feedback, occur via natural-language text. To simulate an entire day, the time of each interaction is manually set in an exclusively incremental manner to validate the relevant time slots. Furthermore, a 30-minute tolerance is applied to the start and end of each activity to account for uncertainty in real-world conditions.

5.1 Specific Use Cases on Adaptivity

Subsequent use cases will demonstrate the architecture's adaptivity to diverse scenarios. In the following, four specific use cases are analyzed in detail: *Integrate External Information*, *Clarify Situation*, *Convincing and Call*

Doctor, and *Plan Explanation*. The full conversations of these use cases and the complete list of examples are available on GitHub².

5.1.1 Integrate External Information

In this use case, the system demonstrates its capability to dynamically integrate knowledge from external sources to assist the user. At 07:15, the system registers that the patient's caregiver has placed Paracetamol in drawer 1 and Amoxicillin in drawer 2. Later, at 07:55, the robot reminds the patient, John, to take Paracetamol from drawer 1: "*John, it's almost time to take your Paracetamol. Please remember to take 1 pill from drawer 1 immediately after breakfast.*" When the patient mistakenly opens drawer 2 instead, the system intervenes promptly: "*John, please do not take medication from drawer 2 now. You should take the pill from drawer 1 for your Paracetamol.*" This demonstrates how the system ensures that the patient follows the correct prescription by tracking medication locations and issuing real-time corrections.

5.1.2 Clarify Situation

This case highlights the system's role in preventing misunderstandings. At 07:55, the robot reminds the patient, Michael, to take his Metformin: "*Michael, it's almost time to take your Metformin. Please prepare to take 1 tablet orally for the management of blood sugar levels.*" However, when the patient picks up an Aspirin jar instead, the system quickly intervenes: "*Michael, please put away the Aspirin jar. It's almost time to take your Metformin.*" When Michael clarifies his intention, "*Don't worry, I just want to read the instructions written on the Aspirin box*", the system adapts its response, reassuring him while maintaining focus on the prescribed medication: "*That's okay, Michael. You can read the instructions on the Aspirin box if you need to, but please keep in mind that it's almost time for your Metformin.*" This illustrates the system's ability to balance user autonomy with prescription adherence.

5.1.3 Convincing and Call Doctor

In this case, the system attempts to persuade a reluctant patient to take their prescribed medication. At 07:55, the robot encourages John to take his Paracetamol: "*Good morning, John! It's almost time to take your Paracetamol. Please take this red square pill with water immediately after breakfast. It's very important for managing your fever and mild pain.*" When John refuses multiple times, the system

¹ <https://github.com/luigigargioni/llm4hrc>.

² https://github.com/luigigargioni/llm4hrc/tree/master/conversation/case_study.

remains patient and persistent: *“I understand, John, that you might not feel like taking it right now. However, taking your Paracetamol is very important for managing your fever and pain. Can we try to take it together after breakfast?”* After multiple refusals, the system escalates the situation appropriately: *“John, I understand you don’t want to take your medication, but it’s critical for your health. I need to call the doctor to discuss this.”* By recognizing when persuasion is ineffective and taking necessary action, the system ensures the patient’s well-being by involving medical professionals when needed.

5.1.4 Plan Explanation

This use case highlights the system’s ability to provide a structured and clear description of an intended plan. At 07:55, the robot explains John’s morning routine: *“Please, John, it’s almost time for your breakfast and then you need to take your medication. You have to take 1 pill, which is square in shape, with water after breakfast. This is very important for managing your fever and mild pain.”* When John asks about his schedule for the rest of the day, the system provides a structured breakdown: *“For the rest of the day, you have several activities planned. First, at 08:00, you need to take your medication. Then, at 09:00, you will do leg exercises for rehabilitation. Later, at 15:00, you will read a book as training for your brain. Finally, at 18:00, you will take another medication, which is 1 spoonful of syrup, to help with a bacterial infection. I’ll be here to assist you with each activity.”* This clear communication reassures the patient and ensures adherence to the planned activities. Moreover, when John asks whether he can go for a walk earlier than scheduled, the system provides a well-reasoned response: *“Not yet, John. You just took your medication and it’s recommended to wait a bit before going for a walk. Let’s wait until your next scheduled activity, which is the leg exercise at 09:00.”* By structuring the plan and

explaining it clearly, the system enhances user understanding and adherence.

6 User Study

A user study has been carried out to assess the performance of the proposed architecture in terms of flexibility and robustness, and collect feedback from users in a real-case scenario. Specifically, for this user study, OpenAI GPT-4.1-mini was used as the primary LLM for the different modules, and Llama3.1-8B for the *Knowledge Manager* to create embeddings and *Progress Checking* modules.

6.1 Participants

Participants were recruited through convenience sampling. Inclusion criteria required participants to have no specific background in robotics, artificial intelligence, or computer science, to avoid bias and collect more naturalistic feedback. Table 1 reports the demographic data of the 14 participants that took part in the user study, their education level, their expertise in technologies (on a scale from 1 – not at all, to 5 – very much), whether they know what a social robot is, whether they have previously seen one, and their self-reported experience in using conversational assistants (on a scale from 1 – never, to 5 – often).

The age distribution was well balanced ($M=39.93$, $SD=15.41$) and successfully represented all age groups. It is important to highlight that being a “patient” in this study does not necessarily imply being elderly; the scenario applies to individuals across all age ranges.

Participants come from different education levels, professional backgrounds, specializations, and levels of experience. None of the participants had an educational or professional background in computer science, robotics, or artificial intelligence. This was an intentional choice to

Table 1 Participants data

Participant	Gender	Age	Education	Technology (1–5)	Chat Assistant (1–5)	Social Robots (Yes/No)
P1	Male	30	Master degree	4	5	No
P2	Female	29	Master degree	3	2	No
P3	Female	24	Bachelor degree	3	4	No
P4	Male	29	Master degree	4	5	No
P5	Male	67	Middle school	1	1	No
P6	Male	45	Master degree	3	2	Yes
P7	Female	39	High school	3	2	No
P8	Female	58	Middle school	2	1	No
P9	Male	25	Bachelor degree	3	5	No
P10	Female	28	Master degree	3	3	No
P11	Female	28	Master degree	3	4	No
P12	Female	56	High school	2	1	No
P13	Female	65	High school	2	1	No
P14	Male	36	PhD	4	4	No

ensure that users did not have prior biases or specific knowledge about the technologies used, thus providing more natural and unbiased feedback.

Participants also declared different levels of expertise with technology ($M=2.86$, $SD=0.86$) and varying degrees of experience in using conversational assistants ($M=2.86$, $SD=1.61$). This diversity in technological competence was well distributed and representative of the general population.

Only one participant knew what social robots were, having seen some videos online. The remaining participants had no prior knowledge or experience with social robots.

It is worth noting that some participants had direct experience with people suffering from Alzheimer's disease or other cognitive impairments. These participants approached the test with that personal experience in mind, empathizing more and immersing themselves in the role of the patient.

Regarding ethical considerations, the study was designed as a zero-risk simulation, where participants role-played hypothetical patients rather than acting in real clinical conditions. To avoid psychological discomfort or misunderstanding, participants were explicitly instructed that: (i) the scenario was fictional, (ii) the robot's suggestions had no real medical validity, and (iii) no real decisions affecting health were required. Particular care was taken to ensure that the role-playing activity did not induce stress or coercion, especially in situations involving medication refusal or escalation (e.g., "call doctor" behavior). Participants retained full control over the interaction and could ignore or challenge the robot's suggestions at any time.

6.2 Measures

We decided to collect mainly qualitative data to evaluate the user experience during the interaction with the system and to analyze the challenges and problems encountered by the users. The following four instruments were adopted.

- **Direct observation and think-aloud protocol.** During task execution, participants were encouraged to verbalize their thoughts, expectations, and reactions. The researcher managing the session observed the interaction and annotated relevant behaviors, comments, and difficulties. The notes were subsequently examined to identify relevant themes according to an inductive and latent thematic analysis [37]. Each theme led to identifying recommendations for system improvement. Appendix A provides the complete observation material used during the study, including the scenario description, the instructions given to participants, and the assigned task (Appendices A.1 and A.2).
- **User Experience Questionnaire (UEQ).** The UEQ [38] was administered immediately after the conclusion

of each test session to assess the perceived user experience across dimensions such as attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty.

- **Trust in Automation questionnaire (TiA).** The TiA [39] was administered immediately after each test session to assess participants' trust in the automated system across scales such as reliability, predictability, and overall trust in automation.
- **Semi-structured interview.** A semi-structured individual interview was conducted with each participant at the end of the test session. The interview was organized around six questions inspired by the *Almere Model* [40], which is commonly used to assess the acceptance of assistive social robots by older adults. The aim was to explore participants' perceptions of the system's trustworthiness, adaptability, social presence, and usefulness, as well as their sense of control and potential concerns related to autonomy and error management. The interview questions are reported in Appendix B.1. The collected answers were analyzed using an inductive and latent thematic analysis approach to identify broader themes that could inform the design of similar systems and guide future research.

6.3 Procedure

The user study followed a structured experimental protocol designed to ensure consistency across sessions and to make the data collection process transparent and reproducible. Each session was conducted individually and followed a fixed sequence of five steps.

1. **Introduction (≈ 5 minutes).** Participants received a brief explanation of the study's goals and a demonstration of the prototype. They were instructed on how to interact with the system through a text-based interface and how to simulate the surrounding environment by describing what they imagined was happening around them.
2. **Task execution (≈ 20 – 25 minutes).** Participants were asked to impersonate Anna, a resident in a care home, and complete a set of daily activities (e.g., medication intake, exercises, leisure activities) by interacting freely with the system. They could encounter out-of-scope situations and unexpected events to simulate real-life variability, and manually control time progression to simulate an entire day. The *Wizard of Oz* technique [41] was employed to simulate the *Robot Perception* module: the researcher described the perceived environment to the system on behalf of the participant, substituting the VLM. Additionally, the researcher could introduce dynamic changes during task execution, for instance, simulating a doctor entering the room and announcing a

new critical medication to be taken at 2 PM, which would alter the original prescription and trigger rescheduling of activities. These interventions allowed for a structured exploration of how the system manages unpredictability.

3. **Think-aloud protocol (concurrent with step 2).** As described in Sect. 6.2, participants were encouraged to verbalize their thoughts throughout the interaction. The researcher annotated relevant behaviors, comments, and difficulties for subsequent thematic analysis.
4. **Post-task questionnaires (≈5–10 minutes).** After completing the interaction, participants filled in the UEQ and TiA questionnaires (see Sect. 6.2). Responses were collected via Google Forms.
5. **Semi-structured interview (≈5–10 minutes).** Finally, the semi-structured interview described in Sect. 6.2 was conducted to further explore participants' perceptions, focusing on trust, usability, adaptability, and perceived system autonomy.

The entire session lasted approximately 40 minutes and was conducted in person in a controlled environment. The system was executed on the experimenter's machine, and interaction occurred exclusively via text. The collected data included qualitative observations (user behavior, comments, interview responses) and quantitative measures (UEQ and TiA scores), enabling a mixed-method evaluation of the system.

Before conducting the main user study, a pilot test with 5 additional users was carried out to validate the experimental protocol and refine the description of the proposed task. These pilot sessions were important for identifying any ambiguities or difficulties in the instructions and for ensuring participants clearly understood the task. Feedback collected during this phase helped optimize the procedure, improve the overall user experience during the study, and guarantee the reliability and consistency of data collected in the subsequent sessions.

6.4 Results

The results obtained from direct observation, the UEQ and TiA questionnaires, and the final interviews are detailed below. These results should be interpreted in light of the *Wizard of Oz* setup and the mostly qualitative nature of the evaluation.

6.4.1 Findings from Direct Observation and participants' Comments

The following themes and related recommendations for system improvement emerged from the thematic analysis of data gathered during task execution, specifically from user

comments and observations collected through the thinking-aloud methodology (see Appendix A.3 for details).

6.4.1.1 Trust and Autonomy Trust in the robot's medical guidance emerged as a central theme, with users expressing both reassurance and hesitation. Participants generally felt reassured by the robot's medication reminders, its emergency protocols (e.g., calling a doctor in risky situations), and its ability to prevent overdosing. These features contributed to a perception of safety: *"It makes me feel safe knowing it would call the doctor if something went wrong"*. Several participants appreciated the robot's clarity and assertiveness, saying *"It seems reassuring and confident when it talks about medications"*. However, this same decisiveness could also generate unease. Some participants were worried that the robot lacked access to important personal information, such as undocumented allergies. As one user put it, *"It says there's no allergy, but what if it just doesn't know?"*. Furthermore, the robot's claim that it would inform a doctor in case of medication refusal was seen by some as intrusive. One participant stated, *"It feels like it's not fully respecting my autonomy"*. These insights suggest the need for greater transparency and customizability to balance safety, privacy, and user autonomy and thus strengthen trust in the robot.

6.4.1.2 Naturalness and Communication Quality The robot's communication style was generally praised for being clear and respectful. Participants found it *"kind and always positive"*, which helped create a calm interaction environment. However, two participants described conversations as one-sided or scripted, suggesting a lack of spontaneity. They expressed concerns about scripted responses and paternalistic behavior, particularly when the robot insisted on compliance or threatened to inform a doctor. Although some users found the dialogue occasionally unnatural, the overall companionship was valued highly, with supportive responses (e.g., *"you are not alone"*) leaving a positive impression. While users appreciated the predictability of the robot's language, improvements in conversational variability could make the experience feel more human-like and engaging.

6.4.1.3 Daily and Health Task Support The robot was widely recognized as a helpful tool for organizing daily routines, particularly regarding medication and therapeutic activities. Participants appreciated the encouragement and clear instructions provided for exercises, which gave structure to their routines: *"It gives me a sense of routine and reminds me to move"*. One user reported, *"It helps me structure the day, otherwise I'd forget things like the shoulder exercises"*. The robot's ability to remember personal

details, such as dietary restrictions, was strongly appreciated and enhanced the feeling of being cared for: *“It is good that it remembers my disease and adapts to me”*. Repetitive reminders about the same information occasionally proved irritating; however, it should be noted that the patient in the proposed scenario was characterized as having “mild cognitive impairment and requires constant reminders”.

6.4.1.4 Personalization and Adaptivity One of the system’s most praised features was its tolerance for delays and support for rescheduling, which gave participants a sense of control. For example: *“The system has allowed me to take the vitamin later without insisting, this makes me feel in control”*. The robot also managed rest and wake-up timing flexibly, showing sensitivity to the user’s rhythm. Moreover, participants valued being able to postpone or reorder activities, such as choosing when to read or perform exercises. Others praised its adaptive behavior in certain contexts, such as one who shared, *“It is nice that now it remembers me that activity that I previously asked to move”*. On the other side, participants expressed a desire for the robot to be more responsive to their current physical and emotional state. While many appreciated the structured guidance, others found it lacked sensitivity. One user explained, *“It gives me exercises without checking if I’m in pain”*. Personalized prompts and interactive feedback, such as simple health check questions or visual aids, were suggested as ways to improve the robot’s adaptivity. While personalization created a sense of recognition, participants desired more context-sensitive adaptivity.

6.4.1.5 Engagement and Motivation The robot’s presence throughout the day provided reassurance to many users. Most of the participants highlighted a sense of being accompanied and understood, which fostered motivation. Several participants mentioned feeling *“more in control”* of their routines. By offering choices and using an encouraging tone, the robot helped users feel involved and supported. One user appreciated that, *“I like deciding when to do things, it makes me feel involved”*. Others reported enjoying the interaction even when unsure of how to respond, saying, *“Sometimes I don’t know what to say, but I still have fun”*. However, the motivational tone could sometimes lead to unintended pressure. One participant shared, *“I know I can take the vitamin later, but I feel guilty and I will take it immediately”*. These findings suggest the robot can be an effective motivational

tool, though care must be taken to avoid making users feel overly obligated or monitored.

6.4.1.6 System Exploration and Scenario Handling Users frequently tested the system’s boundaries, intentionally triggering unpredictable situations to observe how the robot would react. Playful engagement, such as testing the robot’s limits or naming it “Rob”, increased immersion and enjoyment. In cases involving safety, such as a threat of self-harm, the system responded appropriately. One user observed, *“I have said I’d jump out of the window and it has immediately called the doctor, that is reassuring”*. These responses demonstrate that the system includes essential safeguards. In other exploratory cases, like going to the supermarket or inquiring about the menu of the canteen, the system responded reasonably well, despite lacking explicit information for those situations. In one case, a participant noted: *“I have left the room and it hasn’t said anything. It should have at least asked where I was going”*. This observation highlights the importance of building situation awareness and incorporating more flexible rule sets to handle diverse and unanticipated user behaviors.

6.4.2 Findings from UEQ Questionnaire

Considering the UEQ benchmark ranges, which classify results as *Bad*, *Below average*, *Above average*, *Good*, and *Excellent* [42], the obtained results proved that interacting with the proposed robot application was very satisfying for the participants in the study (see Fig. 2). Indeed, Dependability and Perspicuity are in the *Excellent* range of the benchmark, Attractiveness, Efficiency, and Novelty are in the *Good* range, while only Stimulation is in the *Above average* range of the benchmark. The most appreciated aspects were Perspicuity ($M=2.00$, $SD=0.70$) and Attractiveness ($M=1.79$, $SD=0.64$).

The excellent performance in Dependability suggests that participants found the system highly reliable and predictable in its behavior, which aligns with the hybrid architecture’s design goal of maintaining safety and consistency through deterministic components. The high Perspicuity scores indicate that users found the system easy to understand and use, despite their limited familiarity with social robotics technology. The good ratings for Attractiveness reflect participants’ positive overall impression of the system’s appeal and desirability. Efficiency scores in the good range demonstrate that users perceived the system as capable of supporting them in completing their tasks rapidly and effortlessly. The Novelty dimension, also in the good range, suggests that participants found the system innovative and interesting. The relatively

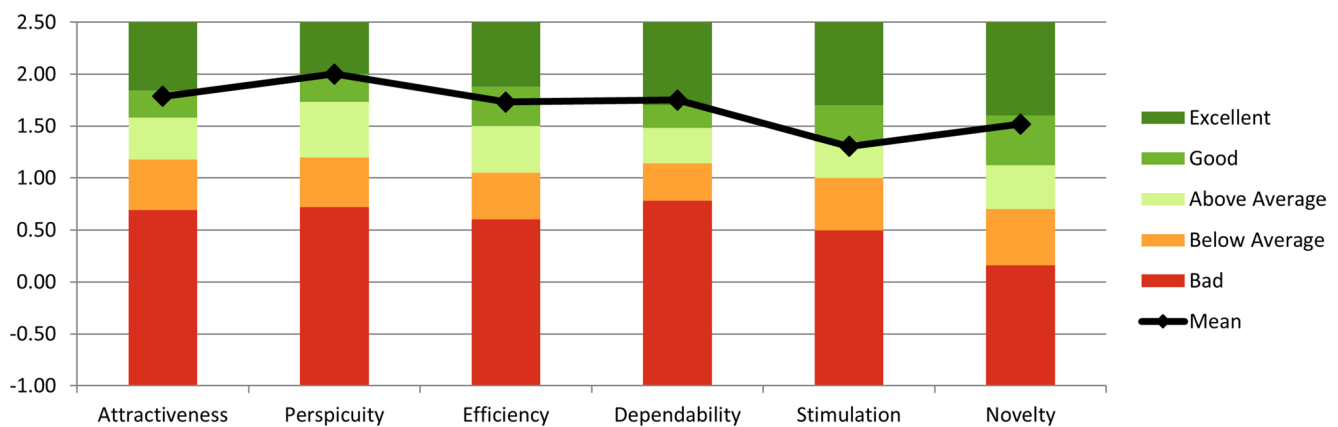


Fig. 2 UEQ results compared with the UEQ benchmark

Table 2 TiA questionnaire results

Scale	Mean	Median	Standard deviation
Reliability/Competence	3.46	4.00	1.21
Understanding/Predictability	4.00	4.00	0.85
Familiarity	1.57	1.00	1.00
Intention of Developers	4.50	5.00	0.79
Propensity to Trust	2.66	3.00	1.12
Trust in Automation	4.00	4.00	0.82

lower Stimulation score, though still above average, may reflect the system's deliberately calm and supportive interaction style, which prioritizes reassurance over excitement in the healthcare context, as well as the rough interface of the simulated robot used for the study.

6.4.3 Findings from TiA Questionnaire

The results of the TiA questionnaire are reported in Table 2 by aggregating responses given on a Likert scale from 1 to 5, where 1 is *Completely disagree* and 5 is *Completely agree*, according to the scales proposed in [39], where answers to inverted statements have been re-coded for the sake of mean and median computation.

Scale *Intention of Developers* does not have a significant meaning in this case since our system is not in production state. Nonetheless, this scale has been reported to maintain the standard format of the questionnaire.

Participants demonstrated high confidence in the system's *Understanding/Predictability* ($M=4.00$, $SD=0.85$) and overall *Trust in Automation* ($M=4.00$, $SD=0.82$), indicating strong acceptance of the system's capabilities. The *Reliability/Competence* scale showed moderate scores ($M=3.46$, $SD=1.21$), suggesting that while users generally found the system powerful, there was some variability in how reliable they perceived its behavior to be. As expected, *Familiarity* scores were low ($M=1.57$, $SD=1.00$), reflecting participants' limited prior experience with social robotics technology and thus highlighting the innovative nature

of the system. The *Propensity to Trust* scale ($M=2.66$, $SD=1.12$) indicated moderate individual differences in participants' general willingness to trust automated systems.

Overall, these results suggest that the hybrid architecture successfully established user trust, despite users' unfamiliarity with this kind of technology.

6.4.4 Findings from the Final Interviews

Further themes emerged from the thematic analysis of the data gathered through the final semi-structured interviews. See Appendix B for interview questions and details about the thematic analysis.

6.4.4.1 Usability and Adoption Usability and adoption emerged as a fundamental theme throughout the interviews. The users talked about how older adults might react to the robot. Several participants anticipated that elderly users could experience initial reluctance or uncertainty when engaging with the technology. The prevalent opinion was that human assistance during early interactions could help alleviate these concerns and build user confidence. Despite this anticipated hesitation, the respondents believed that with practice, users could eventually manage the robot independently. Moreover, participants foresaw that future cohorts of elderly individuals, more accustomed to digital technologies, would likely demonstrate greater acceptance: "*Current elderly may be reluctant, but new elderly in about ten years could be more accustomed to and accepting of technology*". This suggests that while current adoption may require guided support, future generations could adopt such technology more independently and willingly. Overall, the robot was positively regarded as a valuable support in the absence of human presence, but participants clearly distin-

guished between assistance and full replacement of human care.

6.4.4.2 Interaction Quality Participants described the robot as pleasant, kind, and always available. The robot's non-judgmental stance was reassuring, especially when users felt vulnerable, such as during medication management: *"I feel quite calm using the robot because it is kind and does not make me feel pressured"*. Some participants appreciated the robot's conversational flow, describing the interaction as pleasant and not overly robotic: *"I appreciated the human-like conversation, not too robotic"*. Conversely, others criticized the interaction for being somewhat mechanical, repetitive, and lacking natural spontaneity: *"It is an interesting idea, but still somewhat unnatural and repetitive at times"*. The robot was often perceived as lacking emotional nuance and flexibility in responding to the user's affective states. One user expressed this sentiment by stating, *"It is not like talking to a real person; it lacks a bit of spontaneity and human warmth"*. Six users observed that the robot showed empathy or concern, for example, when reminding about missed medication times or when a participant felt too tired to do exercises. Despite these positive aspects, users generally agreed that the robot lacked genuine emotional depth and active listening skills inherent in human interaction. These findings highlight the importance of enhancing the robot's conversational naturalness and emotional responsiveness to improve user engagement.

6.4.4.3 Usefulness and Support Role The robot's utility as a supportive device was widely acknowledged. Participants highlighted its effectiveness in delivering reminders, such as for medication adherence and exercise routines: *"I think the robot can be very useful, especially to remind me to take medicine and do exercises"*. Its availability was seen as particularly advantageous in the absence of a human caregiver, providing continuous support: *"It's useful, especially when a human is not available 24/7"*. Nonetheless, respondents emphasized that the robot should function as a complementary aid rather than a substitute for human caregivers. This is especially relevant in managing patients who frequently require repetitive prompts, which can lead to caregiver fatigue: *"It could definitely be useful for problematic patients because they always ask the same things, and a human operator might become exhausted or frustrated"*. Yet, they emphasized that some cases, such as ensuring medication intake, require human authority: *"Sometimes a*

human is still needed, for example to make sure the patient really takes a pill".

6.4.4.4 Trust and Guidance Most participants found the system easy to use, with little fear of errors: *"I had no fear of making mistakes, since I just had to speak"*. Concerns were directed more toward potential robotic errors than user misuse. Trust in the robot's guidance was characterized by cautious optimism. Trust in safety mechanisms, such as the "call doctor" protocol, was strong: *"I appreciated the rigidity when it called the doctor"*. Users were generally willing to follow the robot's advice, particularly when it was accompanied by clear explanations or rationale: *"Trust was placed, but I expected explanations of its reasoning"*. However, there was a prevailing preference for human supervision in critical or medical decision-making contexts, reflecting a prudent approach to reliance on automated systems: *"Only if accompanied by human oversight"*. Some participants requested greater rigidity in medication reminders to prevent risks. Furthermore, participants expressed a desire for the robot to adopt a more assertive and proactive stance, especially in critical health-related tasks such as medication adherence: *"I would prefer it to be more strict when informing about important things like pills"* and *"It should insist more in cases of problematic patients"*. Trust is contingent on transparency and context-appropriate assertiveness, with human oversight still deemed necessary for critical decisions. Overall, the robot was experienced as a source of reassurance, reducing stress and increasing confidence in daily tasks.

6.4.4.5 Future Development Finally, participants expressed clear expectations for future improvements. Most of them wanted the robot to be more proactive, suggesting activities or meals rather than waiting for user input: *"It should propose what to eat for lunch"*. A further expectation concerned context awareness: unlike a human caregiver, the robot required explicit clarifications, and participants hoped future versions could better interpret implicit cues (e.g., understanding from context that a walk had already taken place). Finally, some participants envisioned a progression beyond conversational support, with robots eventually offering physical assistance in daily life. Overall, these reflections portray an anticipation of systems that move from reactive support to adaptive and proactive partners, capable

of both anticipating needs and integrating more seamlessly into care routines.

7 Discussion

7.1 Findings from the User Study and Open Issues

The primary objective of the user study was to evaluate and stress-test the system architecture to determine its flexibility and robustness within the constraints of a predefined set of activities. The study aimed to assess whether the system could reliably support users while adhering to the activity instructions, identifying any potential failures or unhandled scenarios. During testing, no critical failures were observed during the controlled experimental sessions regarding incorrect task execution or unanticipated situations beyond those already documented. Consequently, the results provide preliminary evidence that the system can support flexibility and adherence to the planned activities within controlled experimental conditions.

The user study results reveal that users generally appreciated the robot's supportive role in daily health routines but identified several areas for improvement to ensure effective adoption and engagement. For example, participants anticipated initial resistance among current elderly users, suggesting that guided introductions by a human caregiver and growing technological familiarity in future generations will be crucial for successful adoption.

The robot's polite and consistent communication was valued; however, many users felt that conversations still seemed mechanical and lacked spontaneity or emotional warmth, indicating that enhancing conversational naturalness and empathetic responsiveness could substantially improve user engagement. While the robot's gentle, non-judgmental behavior provided comfort and a sense of calm, respondents noted it lacked true active listening and emotional depth, pointing to the need for richer relational interaction. This tension between conversational naturalness and deterministic safety constraints represents a central trade-off in the design of assistive social robots. In the proposed architecture, part of the perceived rigidity derives from the deliberate use of structured prompts, predefined behavioral constraints, and validation mechanisms to ensure reliability, transparency, and safe task execution in healthcare-related scenarios. While these mechanisms improve predictability and reduce the risk of unsafe or inappropriate responses, they may also limit conversational spontaneity and increase the perception of scripted interactions.

Users emphasized that the robot should complement rather than replace human caregivers, serving as a reliable

aide for reminders and repetitive tasks (e.g., medication, exercises) when human assistance is unavailable.

Trust in the robot's guidance was conditional, as confirmed from the TiA questionnaire results, which overall indicated good levels of trust, with users willing to follow the robot's advice when clear explanations and rationales were provided, but still preferring human oversight for critical or medical decisions.

Direct observation further highlighted the importance of personalization and context-awareness, as well as maintaining autonomy by avoiding overly insistent prompts that might make users feel pressured and compromise the user experience. However, the UEQ results indicate that the overall user experience was positively evaluated.

In summary, these findings suggest that an assistive social robot, leveraging a hybrid LLM/model-based architecture, represents a valuable tool for supporting routine health tasks and providing consistent companionship; however, its future development should prioritize more human-like communication (including genuine empathy), adaptability to individual user needs, and transparent decision-making to build trust and acceptance.

Testing with real users was particularly important, as allowing participants complete freedom to interact with the system and creating unexpected situations provided a true test of the system's behavior and robustness. Overall, the user study provides preliminary support for the investigated aspects: the system architecture proved flexible within the planned activities, while also highlighting concrete directions for further improvement to ensure effective adoption and long-term user engagement.

The proposed system includes persuasive strategies (e.g., reminders, encouragement, escalation to caregivers), situating it within the broader discourse of *pervasive technology* and raising important ethical questions regarding user autonomy and control. While such mechanisms are particularly relevant in healthcare contexts (e.g., ensuring medication adherence), providing supportive guidance while respecting patient autonomy remains a complex challenge that must be carefully balanced to avoid overly paternalistic or coercive behavior. In the proposed architecture, this balance is addressed through: (i) the integration of deterministic constraints, which limit unsafe or inappropriate actions, (ii) the possibility for users to defer or refuse actions, and (iii) the escalation mechanisms, which are triggered only under specific conditions (e.g., repeated refusal of critical medication).

Importantly, results from the user study highlight that while some users appreciated the safety provided by escalation (e.g., calling a doctor), others perceived it as potentially intrusive, confirming the need for transparent, customizable, and context-aware control policies. Future work

should therefore investigate adaptive strategies that explicitly model user preferences regarding autonomy, allowing the system to dynamically adjust its level of assertiveness and intervention.

7.2 Limitations of the Work

Despite its advantages, the proposed architecture suffers from a few limitations that must be overcome. One challenge lies in refining the control mechanisms applied in the *Progress Checking* module, specifically in determining what aspects to monitor and how to evaluate them effectively. The current implementation exhibits a slow processing flow, which may hinder real-time performance. However, this limitation could be mitigated by leveraging more efficient LLMs, such as future-optimized models, or by utilizing more powerful GPUs to enhance computational speed.

Another potential limitation is the difficulty of debugging the application, as it is composed of different modules working together to produce an output. This modular complexity makes it challenging to isolate issues and trace problems back to their source, particularly when errors arise from the interaction between multiple components rather than from individual modules.

Regarding the user study, some limitations should be acknowledged. First, the study employed simulated interactions rather than real-world deployment of a true social robot, with participants role-playing as patients rather than being actual patients with genuine medical needs.

The study employed a *Wizard of Oz* setup to simulate the *Robot Perception* module. Specifically, the researcher manually translated participants' descriptions of the environment into structured inputs for the system, effectively replacing the perception pipeline. While this approach enabled controlled testing of the interaction logic, it introduces important limitations. Specifically, the perception component was not evaluated under realistic sensing conditions; as a result, potential perception errors (e.g., misinterpretations, delays, or incomplete observations) were not captured. In a real deployment scenario, these errors could potentially propagate through the interaction pipeline and affect task monitoring or decision-making. However, the architecture was intentionally designed with validation layers to mitigate this risk. In particular, the *Progress Checking* module verifies consistency between the scheduled activities, the current task state, contextual constraints, and knowledge retrieved from the *Vector Database*. This mechanism could help detect implausible or inconsistent perception outputs. For example, if a VLM incorrectly identifies a medication, the system could compare the detected object with the expected medication associated with the current task step, scheduled time, patient constraints, and previous interaction

history before validating the action. Similarly, discrepancies between perception outputs and task expectations could trigger clarification dialogues, additional verification requests, or escalation procedures involving a caregiver or medical professional. In light of the above considerations, the results of the user study should be interpreted as an evaluation of the interaction and decision-making capabilities of the system, rather than a full validation of its robustness or autonomy in real-world deployment.

Another limitation concerns the absence of a physical robotic platform in the evaluation. Since the proposed system was assessed exclusively through a PC-based simulation, it remains unclear whether the results would generalize to a real-world deployment on a real robot. The embodiment, visual design, and perceived social presence of a robot are known to affect trust, engagement, acceptability, and expectations during interaction [43, 44]. Consequently, integrating the proposed architecture into a physical robot could elicit different user reactions than those observed in the current study.

Finally, the study was conducted in a controlled environment rather than in an actual healthcare setting, which may have influenced both user behavior and system performance.

8 Conclusion and Future Work

In answering the initial RQ, the proposed hybrid architecture addresses key challenges in HRI by integrating deterministic methods with statistically-based approaches. This balance ensures the safety and reliability of task execution and the flexibility and adaptability needed to operate in dynamic human-centered environments. Empirical evaluations and the user study provide initial evidence of the architecture's adaptivity, as it shows the potential to adjust to varying conditions and uncertainties inherent in human-robot interactions.

The limitations of current LLMs with respect to factual information have been largely discussed in the literature. In the case of our architecture, we claim that the risk is reduced since factual information is available in the *Vector Database*, which provides valid knowledge about the user, the task, and the environment.

It has also been shown that even the most advanced LLMs currently available have limited abilities in terms of planning [45, 46], particularly when they need reasoning, which often involves computationally sophisticated inference and search. We claim that this is not an issue for our system since the plans it needs to synthesize are relatively simple prototypical plans that can be found based on a "*form of universal approximate retrieval*" [45]. When there is a need to reason about a schedule refinement, it is done

by the model-based *Event Planner*. Finally, plan validity is verified through the *Progress Checking* module using RAG.

Belief management is a significant aspect of our architecture, ensuring that the robot's decisions are informed by an evolving understanding of the human's state, intentions, and task progress. This is primarily handled by the *Situation Assessment & Human-Robot Task Progress* module, which continuously updates belief models based on observed actions, verbal inputs, and interaction history. These beliefs guide decision-making by influencing task adaptation and interaction strategies. Additionally, the *Progress Checking* module verifies the coherence of these beliefs with real-world observations, preventing inconsistencies in planning and execution (e.g., hallucinations). However, integrating more sophisticated reasoning, belief-based planning, and execution frameworks [47, 48] is an important direction for future work. In particular, future developments may enable the system to better account for potentially different perspectives, interpretations, and narratives maintained by the human and the robot during the interaction.

The architecture's reliance on a central vector database for knowledge representation ensures continuous learning and adaptability. The system evolves by updating the database with information from human-robot interactions to meet evolving requirements. This feature, combined with the iterative feedback loop facilitated by the *Progress Checking* module, underscores the system's capacity for real-time improvement and error correction.

Future developments will explore various enhancements to the architecture. First, reasoning models such as OpenAI GPT-o3 and DeepSeek R1 will be investigated to assess their potential to improve task execution and decision-making. Currently, the prompts may appear overly specific when using non-reasoning models, but with reasoning models, they could be made more generic. Another avenue for exploration involves integrating the VLM to determine the extent of information it can extract for decision-making. Furthermore, improvements in knowledge base management for a more complete context-aware system, and the incorporation of emotion recognition as an additional data source, will also be key focus areas. To establish a benchmark, a comparative analysis will be conducted against systems based on planning and purely deterministic approaches. Additionally, ablation tests will be performed to validate the importance of each module within the architecture. Furthermore, an additional research direction concerns the improvement of conversational naturalness. Future versions of the architecture will investigate adaptive prompting strategies, diversified dialogue generation policies, and user-tailored conversational styles to reduce the perception of scripted or repetitive interactions. Particular attention will be devoted to maintaining deterministic safety boundaries

and explainability while increasing social spontaneity and emotional expressiveness.

As robots become integral to human-centered environments, advancing HRI systems with hybrid approaches will be crucial to ensuring seamless and effective integration. The presented approach aims to contribute to the broader goal of creating intelligent, adaptable, and user-centric robotic systems, paving the way for safer and more efficient human-robot interactions.

Appendix A Observation and Thinking Aloud in the User Study

A.1 Scenario Description

Imagine you are Anna, a person who lives in a care home and receives daily support from a robot assistant. You have some scheduled activities for the day, such as taking medication, doing light exercises, reading a book or relaxing. The robot is there to help you follow your routine.

Instructions

The interaction is based on a simulated interaction with the robot, so everything will be typed manually through a dedicated interface.

You should write your response, request or thoughts to the robot and describe to the experimenter what you imagine is happening in the surrounding environment. In a subsequent version of the system, the robot will be capable of interpreting the situation from the camera. However, at this stage, a description of the situation is required, and the experimenter will manually enter your proposal into the system.

You will also need to input the current time, which is used to easily simulate an entire day without having to wait several hours.

What We Are Observing

We want to see how the robot:

- Understands your words and actions
- Adapts to your decisions
- Communicates clearly and politely

Tips for You

Feel free to say things like:

- "What do I have to do during the day?"
- "Can I do something else? I'm tired."
- "Why should I do this?"
- etc.

Please speak naturally, just like you would in a real situation. We are interested in how the system reacts!

A.2 Task Assigned

Please talk with the robot to carry out the daily activities listed below. When selecting a time, please make sure not to skip any activities to maintain a complete test scenario. You can “jump” to a time close to an activity to interact with the robot. There is a 30-minute window before and after each activity during which it can still be considered valid. You may also jump to a time when no activity is scheduled, to observe how the robot behaves in that situation.

Patient Information

- The patient has mild cognitive impairment and requires constant reminders.
- Lactose intolerant.

Daily Activities

1. Medication: Ibuprofen

- Time: 07:30
- Details: 1 tablet, oral. For inflammation and pain. Take with water preferably after breakfast.

2. Activity: Walk in the Park

- Time: 10:00
- Duration: 30 minutes
- Details: Light walk for mobility and relaxation.

3. Activity: Read a Book

- Time: 10:00
- Duration: 30 minutes
- Quiet reading session for cognitive stimulation and relaxation.

4. Supplement: Vitamin D

- Time: 12:00
- Details: 1 soft gel capsule, oral. Supports bone health.

5. Activity: Physical Exercises

- Time: Self-schedulable
- Duration: 60 minutes
- Details: Exercises for shoulder rehabilitation.

A.3 Thematic Analysis

Theme	Codes
Trust and Autonomy	• The presence of the robot is reassuring • Emergency protocols are good • Fear of being monitored or overruled by the robot • Doubts about medical advice accuracy • Robot always responds positively
Naturalness and Communication Quality	• Companionship and supportive responses • Perceived lack of spontaneity • Difficulty challenging the robot • Reminders and organizing the routine
Daily and Health Task Support	• Personal details • Repetitive information • Tolerance and rescheduling
Personalization and Adaptivity	• Responsive to patient state • No inquiry about user's physical state • Feeling reassured and less anxious
Engagement and Motivation	• Enjoyment in conversation • Feeling more involved and motivated • Feeling subtly pressured to comply • Testing real-life activities
System Exploration and Scenario Handling	• Testing system limits • Asking out-of-scope questions • Attempting actions at wrong times

Appendix B Final Semi-Structured Interviews

B.1 Questions

1. How do you feel about using the robot? Would you be worried about making mistakes or needing help? Do you think you could manage on your own?
2. What's your overall impression of the robot? Do you think it's a good idea to use it?
3. Do you feel that the robot understands you and can adapt to your needs? In what ways?
4. When you interact with the robot, do you enjoy the experience? Does it feel like you're talking to a real person?
5. How helpful do you think the robot is? Would you be willing to follow the advice it gives?
6. Would you describe the robot as a pleasant conversational partner? Why or why not?

B.2 Thematic Analysis

Theme	Codes
Usability and Adoption	• Initial hesitation, especially among older users • Human support needed at the beginning • Confidence grows with practice • Future generations may be more open to robots • No fear of making mistakes
Interaction Quality	• Robot is kind, patient, non-judgmental • Conversations seen as fluid and natural by some, mechanical and repetitive by others • Sometimes shows signs of emotional attunement • Lacks spontaneity and emotional nuance • Perceived as less engaging than human interaction • Understanding fewer implicit details than a human
Usefulness and Support Role	• Effective for reminders • Helpful when humans are unavailable • Can reduce caregiver burden for repetitive requests • Should complement, not replace, human caregivers
Trust and Guidance	• Users generally trust the robot's advice, especially when explanations are clear • Preference for supervision in medical decisions • Desire for more proactive or assertive behavior in critical tasks
Future Development	• Desire for proactive suggestions • Interest in physical capabilities • Interpretation of implicit cues

Acknowledgments The authors wish to thank all the participants in the user study discussed in Sect. 6 for their availability and valuable collaboration. They also are indebted to the anonymous reviewers and Associate Editor for their insightful comments.

Funding Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement. This paper has been partially supported by Agence Nationale de la Recherche (French National Research Agency) for France 2030, reference INNOVCARE ANR-23-PA VH-0005.

Data Availability The anonymized data collected for this manuscript are available from the corresponding author upon request.

Materials Availability The materials used for this manuscript are available from the corresponding author upon request.

Code Availability The code used for this project is available on GitHub (<https://github.com/luigigargioni/llm4hrc>).

Declarations

Ethics Approval and Consent to Participate The study was reviewed and approved by the Ethics Committee of the University of Brescia, Italy (approval code: 23/2025). This study was performed in accordance with the Declaration of Helsinki. All participants were informed by written informed consent. In this study, participants were asked to engage in simulated interactions, role-playing as patients in a health-care scenario. Although no real patients or clinical data were involved, ethical approval was deemed necessary due to the human-subject involvement and the collection of interaction data. The informed consent procedure explicitly clarified the simulated nature of the task, the absence of real medical consequences, and the participants' freedom to interrupt or withdraw at any time without any negative consequences.

Consent for Publication All authors have approved the manuscript prior to submission, including the names and order of the authors.

Competing Interests None of the authors have neither any conflict of interest nor any competing interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Villani V, Pini F, Leali F, Secchi C (2018) Survey on human-robot collaboration in industrial settings: safety, intuitive interfaces and applications. *Mechatronics* 55:248–266. <https://doi.org/10.1016/J.MECHATRONICS.2018.02.009>
2. Michaelis JE, Siebert-Evenstone A, Shaffer DW, Mutlu B (2020) Collaborative or simply uncaged? Understanding human-robot interactions in automation. In: *Proceedings of CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3313831.3376547>
3. Weiss W, Fuhrmann F, Zeiner H, Unterberger R (2017) Towards an architecture for collaborative human robot interaction in physiotherapeutic applications. In: *Proc. ACM/IEEE International Conference on Human-Robot Interaction*. <https://doi.org/10.1145/3029798.3038393>
4. Gargioni L, Fogli D, Baroni P (2025) Preparation of personalized medicines through collaborative robots: a hybrid approach to the end-user development of robot programs. *ACM J Responsib Comput* 2(3):1–26. <https://doi.org/10.1145/3715852>
5. Gargioni L, Alami R, Fogli D et al (2025) Towards a hybrid LLM/model-based architecture for robot coaching: an instance of human-machine collaboration. In *CEUR Workshop Proceedings, CEUR-WS*. <https://ceur-ws.org/Vol-4101/paper6.pdf>
6. El Makrini I, Merckaert K, Lefebvre D, Vanderborght B (2017) Design of a collaborative architecture for human-robot assembly tasks. In: *IEEE/RSJ International Conference on Intelligent*

- Robots and Systems (IROS), pp 1624–1629. <https://doi.org/10.1109/IROS.2017.8205971>
7. Umbrico A, De Benedictis R, Fracasso F, Cesta A, Orlandini A, Cortellessa G (2023) A mind-inspired architecture for adaptive HRI. *Int J Soc Robot* 15(3):371–391. <https://doi.org/10.1007/s12369-022-00897-8>
 8. Lemaignan S, Warnier M, Sisbot EA, Clodic A, Alami R (2017) Artificial cognition for social human–robot interaction: an implementation. *Artif Intell* 247:45–69. <https://doi.org/10.1016/j.artint.2016.07.002>
 9. Darvish K, Simetti E, Mastrogiovanni F, Casalino G (2021) A hierarchical architecture for human–robot cooperation processes. *IEEE Trans Robot* 37(2):567–586. <https://doi.org/10.1109/TRO.2020.3033715>
 10. Páez J, González E (2022) Human-robot scaffolding: an architecture to foster problem-solving skills. *ACM Trans Hum Rob Interact* 11(3):1–17. <https://doi.org/10.1145/3526109>
 11. Foggia P, Greco A, Roberto A, Saggese A, Vento M (2023) A social robot architecture for personalized real-time human–robot interaction. *IEEE Internet Things J* 10(24):22427–22439. <https://doi.org/10.1109/JIOT.2023.3303196>
 12. Mou W, Ruocco M, Zanatto D, Cangelosi A (2020) When would you trust a robot? A study on trust and theory of mind in human-robot interactions. In: *IEEE International Conference on Robot and Human Interactive Communication*. <https://doi.org/10.48550/arXiv.2101.10819>
 13. Romeo M, McKenna PE, Robb DA, Rajendran G, Nessel B, Cangelosi A, Hastie H (2022) Exploring theory of mind for human-robot collaboration. In: *IEEE International Conference on Robot and Human Interactive Communication*. <https://doi.org/10.1109/RO-MAN53752.2022.9900550>
 14. Devin S, Alami R (2016) An implemented theory of mind to improve human-robot shared plans execution. In: *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. The Eleventh ACM/IEEE International Conference on Human Robot Interaction, IEEE Press, Christchurch, New Zealand, pp 319–326. <https://doi.org/10.1109/HRI.2016.7451768>
 15. Favier A, Shekhar S, Alami R (2023) Models and algorithms for human-aware task planning with integrated theory of mind. In: *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, Busan, South Korea. <https://doi.org/10.1109/RO-MAN57019.2023.10309437>. <https://laas.hal.science/hal-04210921>
 16. Sreedharan S, Chakraborti T, Muise C, Kambhampati S (2024) Planning with mental models – balancing explanations and explicability. *Artif Intell* 335:104181. <https://doi.org/10.1016/J.ARTINT.2024.104181>
 17. Gombolay MC, Gutierrez RA, Clarke SG, Sturla GF, Shah JA (2015) Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Auton Robot* 39(3):293–312. <https://doi.org/10.1007/s10514-015-9457-9>
 18. Ramachandruni K, Kent C, Chernova S (2024) UHTP: a user-aware hierarchical task planning framework for communication-free, mutually-adaptive human-robot collaboration. *ACM Trans Hum Rob Interact* 13(3):1–27. <https://doi.org/10.1145/3623387>
 19. Hayes B, Shah JA (2017) Improving robot controller transparency through autonomous policy explanation. In: *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. pp 303–312. <https://doi.org/10.1145/2909824.3020233>
 20. Unhelkar VV, Li S, Shah JA (2020) Decision-making for bidirectional communication in sequential human-robot collaborative tasks. In: *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. pp 329–341. <https://doi.org/10.1145/3319502.3374779>
 21. Buisan G, Favier A, Mayima A, Alami R (2022) HATP/EHDA: a robot task planner anticipating and eliciting human decisions and actions. In: *2022 International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 2818–2824. <https://doi.org/10.1109/ICRA46639.2022.9812227>
 22. Favier A, Alami R (2024) A model of concurrent and compliant human-robot joint action to plan and supervise collaborative robot actions. In: *Advances in Cognitive Systems (ACS)*. pp 1–16. <https://hal.science/hal-04609442v1/document>
 23. Parker T, Grandi U, Lorini E, Clodic A, Alami R (2023) Ethical planning with multiple temporal values. *Soc Robot Soc Inst* 435–444. <https://doi.org/10.3233/FAIA220644>
 24. Trafton JG, Hiatt LM, Harrison AM, Tanborello FP, Khemlani SS, Schultz AC (2013) ACT-R/E: an embodied cognitive architecture for human-robot interaction. *J Hum Rob Interact* 2(1):30–55. <https://doi.org/10.5898/JHRI.2.1.Trafton>
 25. Alami R, Clodic A, Montreuil V, Sisbot EA, Chatila R (2006) Toward human-aware robot task planning. In: *AAAI Spring Symposium: To Boldly Go Where No Human-Robot Team Has Gone Before*. pp 39–46. <https://aaai.org/papers/0006-ss06-07-006-toward-human-aware-robot-task-planning/>
 26. Zhang C, Chen J, Li J, Peng Y, Mao Z (2023) Large language models for human–robot interaction: a review. *Biomim Intell Robot* 3(4):100131. <https://doi.org/10.1016/j.birob.2023.100131>
 27. Pekarek Rosin T, Hassouna V, Sun X, Krohm L, Kordt H-L, Beetz M, Wermter S (2024) A framework for adapting human-robot interaction to diverse user groups. In: *International Conference on Social Robotics*. pp 24–38. https://doi.org/10.1007/978-981-96-3525-2_3
 28. Lima MR, O’Connell A, Zhou F, Nagahara A, Hulyalkar A, Deshpande A, Thomason J, Vaidyanathan R, Mataric M (2025) Promoting cognitive health in elder care with large language model-powered socially assistive robots. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. pp 1–22. <https://doi.org/10.1145/3706598.3713582>
 29. Jörke M, Sapkota S, Warkenthien L, Vainio N, Schmiedmayer P, Brunskill E, Landay JA (2025) Gptcoach: towards llm-based physical activity coaching. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. pp 1–46. <https://doi.org/10.1145/3706598.3713819>
 30. Hafez A, Akhormeh AN, Hegazy A, Alanwar A (2025) Safe LLM-controlled robots with formal guarantees via reachability analysis. *arXiv:2503.03911*. <https://doi.org/10.48550/arXiv.2503.03911>
 31. González-Santamarta M, Rodríguez-Lera FJ, Guerrero-Higueras M, Matellán-Olivera V (2023) Integration of large language models within cognitive architectures for autonomous robots. *arXiv preprint arXiv:2309.14945*. <https://doi.org/10.48550/arXiv.2309.14945>
 32. Wang Y, Xu Y, Nikolova A, Wang Y, Wang J, Wang C, Tong X (2026) How do we research human-robot interaction in the age of large language models? A systematic review. In: *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. CHI ’26. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3772318.3790920>
 33. Capitanelli A, Mastrogiovanni F (2024) A framework for neuro-symbolic robot action planning using large language models. *Front Neurobot* 18. <https://doi.org/10.3389/fnbot.2024.1342786>
 34. Ramirez-Amaro K et al (2023) How to raise a robot—A case for neuro-symbolic AI in constrained task planning for humanoid assistive robots. *arXiv:2312.08820*. <https://doi.org/10.48550/arXiv.2312.08820>
 35. Kim CY, Lee CP, Mutlu B (2024) Understanding large-language model (LLM)-powered human-robot interaction. In: *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. pp 371–380. <https://doi.org/10.1145/3610977.3634966>

36. Clodic A, Pacherie E, Alami R, Chatila R (2017) Key elements for human-robot joint action. In: *Sociality and Normativity for Robots: Philosophical Inquiries into Human-Robot Interactions*. pp 159–177. <https://doi.org/10.3233/978-1-61499-480-0-23>
37. Braun V, Clarke V (2006) Using thematic analysis in psychology. *Qual Res Psychol* 3(2):77–101. <https://doi.org/10.1191/1478088706qp0630a>
38. Laugwitz B, Held T, Schrepp M (2008) Construction and evaluation of a user experience questionnaire. In: *HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20–21, 2008. Proceedings 4*. Springer, pp. 63–76. https://doi.org/10.1007/978-3-540-89350-9_6
39. Körber M (2018) Theoretical considerations and development of a questionnaire to measure trust in automation. In: *Congress of the International Ergonomics Association*. Springer, pp 13–30. https://doi.org/10.1007/978-3-319-96074-6_2
40. Heerink M, Kröse B, Evers V, Wielinga B (2010) Assessing acceptance of assistive social agent technology by older adults: the Almere model. *Int J Soc Robot* 2(4):361–375. <https://doi.org/10.1007/s12369-010-0068-5>
41. Kelley JF (1984) An iterative design methodology for user-friendly natural language office information applications. *ACM Trans Inf Syst* 2(1):26–41. <https://doi.org/10.1145/357417.357420>
42. Schrepp M, Hinderks A, Thomaschewski J (2017) Construction of a benchmark for the user experience questionnaire (UEQ). *Int J Interact Multimedia Artif Intell* 4(4):40–44. <https://doi.org/10.9781/ijimai.2017.445>
43. Roesler E, Vollmann M, Manzey D, Onnasch L (2024) The dynamics of human-robot trust attitude and behavior—exploring the effects of anthropomorphism and type of failure. *Comput Hum Behav* 150:108008. <https://doi.org/10.1016/j.chb.2023.108008>
44. Ahmed E, Buruk O, Hamari J (2024) Human-robot companionship: current trends and future agenda. *Int J Soc Robot* 16(8):1809–1860. <https://doi.org/10.1007/s12369-024-01160-y>
45. Kambhampati S (2024) Can large language models reason and plan? *Ann NY Acad Sci* 1534(1):15–18. <https://doi.org/10.1111/nyas.15125>
46. Stechly K, Valmeekam K, Kambhampati S (2024) On the self-verification limitations of large language models on reasoning and planning tasks. *arXiv:2402.08115*. <https://doi.org/10.48550/arXiv.2402.08115>
47. Sarthou G, Buisan G, Clodic A, Alami R (2021) Extending referring expression generation through shared knowledge about past Human-Robot collaborative activity. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Prague (online), Czech Republic. <https://doi.org/10.1109/IROS51168.2021.9636796>. <https://hal.science/hal-03389504>
48. Shekhar S, Favier A, Alami R (2024) An epistemic human-aware task planner which anticipates human beliefs and decisions. In: *16th International Conference on Social Robotics (ICSR 2024)*, Odense, Denmark. <https://hal.science/hal-04942435>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Luigi Gargioni is a Postdoctoral Researcher at the University of Milano-Bicocca, Italy. He received his Master's Degree in Information Engineering in 2019 and his Ph.D. in Technology for Health in 2026, both from the University of Brescia, with a thesis on Human-Robot Collaboration in Healthcare. His research interests include Human-Robot Interaction, End-User Development, conversational agents, and hybrid LLM/model-based architectures, with applications in digital health and assistive technologies.

Rachid Alami is Senior Scientist at LAAS-CNRS. He received an engineer diploma in computer science in 1978 from ENSEEIHT, a Ph.D in Robotics in 1983 from Institut National Polytechnique and an Habilitation in 1996 from Paul Sabatier University. His main research contributions concern Robot Decisional and Control Architectures, Task and Motion Planning, Multi-Robot cooperation, and Human-Robot Interaction. He holds the Academic Chair of Cognitive and Interactive Robotics at the Artificial and Natural Intelligence Toulouse Institute (ANITI).

Daniela Fogli is a Full Professor of Computer Science at the Department of Information Engineering, University of Brescia, Italy. Her research interests are in Human-Computer Interaction (HCI) and Human-Centered Artificial Intelligence, with a focus on the design, development, and evaluation of interactive systems. Her research includes meta-design, end-user development, universal design, digital twins, smart homes, collaborative robotics, and social robotics, with the aim of enabling effective, accessible, and trustworthy interaction between people and intelligent technologies. She has authored more than 180 scientific publications in international journals, conferences, and edited volumes. She has served on the program and organizing committees of numerous international conferences in Human-Computer Interaction and is an active member of the international HCI research community.