2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

GPTAlly: A Safety-Oriented System for Human-Robot Collaboration Based on Foundation Models

Brieuc Bastin[1], Shoichi Hasegawa[2], Jorge Solis[3], Renaud Ronsse[1], Benoit Macq[1], Lotfi El Hafi[2], Gustavo Alfonso Garcia Ricardez[2,*], and Tadahiro Taniguchi[2,4]

***Abstract*— As robots increasingly integrate into the workplace, Human-Robot Collaboration (HRC) has become increasingly important. However, most HRC solutions are based on pre-programmed tasks and use fixed safety parameters, which keeps humans out of the loop. To overcome this, HRC solutions that can easily adapt to human preferences during the operation as well as their safety precautions considering the familiarity with robots are necessary. In this paper, we introduce GPTAlly, a novel safety-oriented system for HRC that leverages the emerging capabilities of Large Language Models (LLMs). GPTAlly uses LLMs to 1) infer users' subjective safety perceptions to modify the parameters of a Safety Index algorithm; 2) decide on subsequent actions when the robot stops to prevent unwanted collisions; and 3) re-shape the robot arm trajectories based on user instructions. We subjectively evaluate the robot's behavior by comparing the safety perception of GPT-4 to the participants. We also evaluate the accuracy of natural language-based robot programming of decision-making requests. The results show that GPTAlly infers safety perception similarly to humans, and achieves an average of 80% of accuracy in decision-making, with few instances under 50%. Code available at: https://axtiop.github.io/GPTAlly**

I. Introduction

Our society is aiming for Society 5.0[†], a human-centric society that emphasizes workplace quality of life, unlike Society 4.0, which prioritizes productivity often at the cost of well-being [1]. In Society 5.0, robots and humans collaborate with AI assistance to complete tasks, reducing strain and fostering a healthier work environment. Achieving this requires innovative tools such as Large Language Models (LLMs) and Vision Language Models (VLMs), which exhibit emergent properties such as natural language understanding [2], context awareness [3], and zero-shot prompting [4].

This work was supported by the Japan Science and Technology Agency (JST), Moonshot Research & Development Program, Grant Number JPMJMS2011.

[1]Brieuc Bastin, Benoit Macq, and Renaud Ronsse are with Université catholique de Louvain (UCLouvain); 1 Place de l'Université, Louvain-la-Neuve 1348, Belgium. brieuc.bastin@student.uclouvain.be, {benoit.macq, renaud.ronsse}@uclouvain.be

[2]Shoichi Hasegawa, Lotfi El Hafi, Gustavo Alfonso Garcia Ricardez, and Tadahiro Taniguchi are with Ritsumeikan University; 1-1-1 Noji-Higashi, Kusatsu, Shiga 525-8577, Japan. {hasegawa.shoichi, lotfi.elhafi, garcia-g, taniguchi}@em.ci.ritsumei.ac.jp

[3]Jorge Solis is with Karlstad University; 2 Universitetsgatan, Karlstad 651 88, Sweden. jorge.solis@kau.se

[4]Tadahiro Taniguchi is with Kyoto University; Yoshida-Honmachi, Sakyo, Kyoto 606-8501, Japan. taniguchi@i.kyoto-u.ac.jp

*Corresponding author.

[†]As defined by the Japan Cabinet Office: https://www8.cao.go.jp/cstp/english/society5_0/

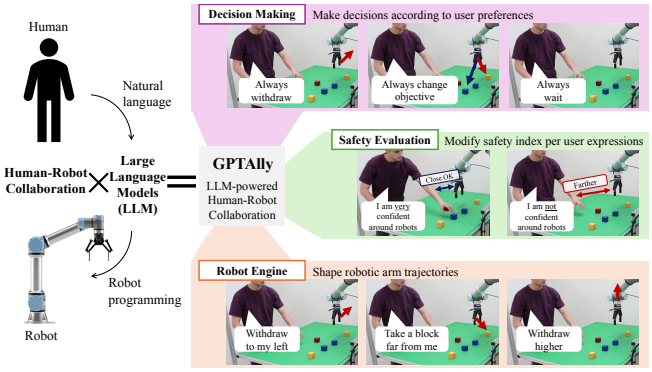


Fig. 1. Representation of human-robot collaboration performed with GPTAlly with adapted stopping distance based on the user perception of safety, intelligent decision-making, and reactive trajectory shaping.

One challenge with current industrial collaborative robots is their limited human-like understanding, as they are pre-programmed for specific tasks or use supervised learning models, and lack the capability to adapt to new contexts in unstructured environments [3]. Solving this challenge requires leveraging LLMs' emergent properties for contextual understanding and dynamic decision-making while ensuring robustness and real-time processing.

Another challenge is natural language understanding (NLU), which is highly contextual and ambiguous, making it difficult for robots to interpret commands accurately [2]. Robots should recognize and understand natural language commands in a given context and adapt their behavior to user preferences. Integrating foundation models into robotic systems involves overcoming scalability challenges, as the increasing volume of natural language input reduces the accuracy of most models, strains model processing capabilities, and limits generalization across contexts and languages [2], [3], [5].

Additionally, there is a lack of collaborative robot safety evaluations that consider users' subjective perceptions of safety. Users' anxiety about collaborative robots can itself become a source of danger, potentially leading to hesitation, erratic movements, or misjudgments that increase the risk of collisions [6]. Recent safety evaluation algorithms ensure physical Human-Robot Interaction (pHRI) safety but do not consider users' safety perceptions [7]. LLMs can help infer subjective safety perceptions by interpreting human feedback, emotions, and contextual cues.

This paper proposes General Pre-Trained Ally (GPTAlly),

a safe and adaptable system for human-robot collaboration. The main goal is to adapt the robot's behavior based on user safety perception through natural language as shown in Fig. 1. If the user *feels* safe, the robot moves quicker and is less restricted; if the user *feels* unsafe, it slows down and is more restricted. In dangerous situations, the robot stops to ensure safety but can wait, withdraw to a safer location, or change goals autonomously or based on user preferences. A situation is deemed dangerous if the safety evaluation metric proposed in this paper, which incorporates the positions and movements of both the robot and the user, as well as the user's perception of safety, exceeds a defined threshold. The system suggests withdrawal and target positions based on user desires without the use of models trained for narrow tasks.

The main contributions of this work are:

1) Demonstrate how **LLMs can incorporate human perception of safety into safety evaluations** by combining a safety index from the literature with LLMs' interpretation of natural language input.
2) Propose a **streamlined coding paradigm** based on LLMs for decision-making in HRC by using natural language to describe the conditions used to modify the robot behavior.
3) Verify how LLMs can be used to **shape robotic arm trajectories** by suggesting 3D poses based on natural language input and the robot's current state.

II. Related Works

The related works are divided into two categories to contextualize the contributions of this research: the integration of LLMs in robotics and human safety considerations in HRC. These categories emphasize the advancements in LLM-driven decision-making and safety evaluation methods, which underpin the innovations proposed in this work.

Large Language Models and Robotics: Depending on the task the robots must complete, they must be able to achieve a combination of task-specific operations such as code generation [8], open vocabulary object detection [9], high-level task planning [10], or even low-level control [11]. The wide range of tasks requires using different state-of-the-art models, which led to the development of various systems [3].

LATTE [11] uses VLMs and LLMs to modify a specific trajectory. Their general model utilizes pre-trained language models (BERT [12] and CLIP [13]) to encode the user's requests and target objects directly from a text input and scene images. However, GPTAlly leverages the zero-shot capabilities of LLMs to suggest 3D poses, without requiring fine-tuning on additional training data.

AutoTAMP [10] introduces a framework leveraging LLMs to translate language task descriptions into formal task specifications through few-shot in-context learning. Additionally, it employs these LLMs as validators for identifying syntactic and semantic errors via corrective re-prompting. However, in GPTAlly, LLMs are utilized for single-sequence planning without corrective re-prompting, emphasizing real-time solutions rather than multi-task planning capabilities.

Human Safety: There exist many ways to estimate the danger a user faces when collaborating with a robot such as using injury evaluation, hazard analysis, the distance and robot link momentum, or the impact force and impact stress [14], [15], [16], [17]. Based on human pose estimation and robot pose, [18] proposes a safety index to assess human safety by considering the consequences of the most dangerous situation. It links potential danger to potential human injuries by assessing the severity of impact during a collision and foresees human movements to proactively limit the robot's speed, reducing collision likelihood and associated risks. Using the robot state, human state, and worst possible action is not the only method for determining the safety index. However, in GPTAlly, the safety index incorporates a scaling factor suggested by an LLM based on the user perception of safety into an established Safety Index (SI) algorithm suggested by [18].

III. Proposed Method

The overall goal is to realize a safe and adaptable system for human-robot collaboration. Our proposed system consists of three main components, namely, decision-making, robot controller, and safety evaluation, which are intended to be robot agnostic. We first formalize the problem definition, then describe the three main components, and finally, introduce a component needed for integration.

A. Problem Definition

Let L_{in} be the user's natural language input sent to express the user's desires during the collaboration, such as L_{in} = "*I am not confident with robots, please be careful*" or L_{in} = "*When you withdraw, withdraw to the left.*" Let $GPT_f \in [0:2]$ be a scaling factor provided by an LLM (*e.g.*, GPT-4) to infer the user perception of safety with 0 meaning that the user is confident enough to drastically reduce the safety evaluation and 2 meaning that the user wants the robot to stop moving while in the same workspace as the user. Let $P_{\text{human}} \in \mathbb{R}^3$ be the human upper body position the closest to the robot, $P_{\text{robot}} \in \mathbb{R}^3$ be the position of the end effector, and $P_{init} \in \mathbb{R}^3$ be the current objective position where the end effector is moving toward. They are all expressed in the Cartesian frame, the origin of which is at the base of the robotic arm. Finally, $SI \in [0:1]$ is the safety index, with 0 representing the most dangerous situation and 1 the safest.

Let $\mathbf{O} = \{O_1, ..., O_M\}$ be a collection of *M* objects in the environment, each with a corresponding position $P_{O_i} \in \mathbb{R}^3$. The poses are expressed in the Cartesian frame at the base of the robotic arm. Furthermore, $F_{\text{choice}} \in \{1, 2, ..., i\}$ is the index i of the function h_i to be executed by the robot. It is determined by the decision-making module implemented with GPT-4.

Lastly, $P_{obj} \in \mathbb{R}^3$ is the suggested objective position where the end effector should move. $P_{in} \in \mathbb{R}^3$ is the new objective position where the end effector has to move after P_{obj} is processed by the constraint satisfaction module.

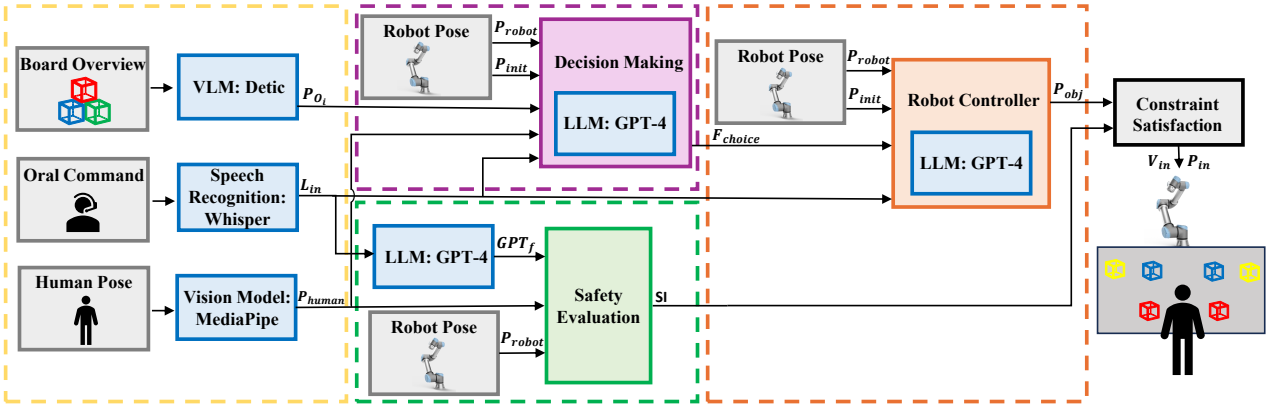


Fig. 2. System architecture: in yellow, the multimodal input module is composed mostly of pre-trained models. In green, the safety evaluation module consists of an SI algorithm and a GPT factor (GPT_f) computed by GPT-4. In purple, the decision-making module is designed with GPT-4. In orange, the robot functions and a constraint satisfaction sub-module are implemented.

$V_{in} \in [0:1]$ is a scaling factor computed by the constraint satisfaction module to scale the speed based on the *SI*. The positions are represented in the Cartesian frame at the base of the robotic arm.

$$\begin{cases} SI = f(L_{in}, P_{init}, P_{\text{human}}, P_{\text{robot}}), \\ F_{\text{choice}} = g(L_{in}, P_{init}, P_{O_i}, P_{\text{human}}, P_{\text{robot}}), \\ P_{obj} = h(L_{in}, P_{init}, P_{O_i}, P_{\text{human}}, P_{\text{robot}}, F_{\text{choice}}). \end{cases} \tag{1}$$

The first goal is to design a function f that evaluates the safety (SI) based on the robot's position, the current objective position, the human limbs' position, and the oral commands. The second goal is to design a function g that maps the current state of the robot to a specific function F_{choice}. Lastly, the function h implements the functions with the index F_{choice} and suggests the position to which the robot should go.

B. Proposed System Architecture

The proposed system represented in Fig. 2 uses the functions f, g, and h from Equation 1 to provide a safe and interactive system. These functions are non-trivial and uncertain as they integrate data from diverse modalities, and the solution space has multiple solutions that fulfill the user's semantic preferences. The model architecture is divided into an input module (in yellow), three function modules (one per function), and a constraint satisfaction module:

Multimodal perception: The multimodal perception module collects two distinct types of data: proprioceptive and exteroceptive data. Proprioceptive data, inherent to the robot system, encompasses the positioning of each joint within the robot and the Cartesian coordinates of the robot's end effector, denoted as P_{robot}. This data type comprises sensory inputs obtained from oral commands and visual inputs. Initially, user commands are captured by a microphone, denoted as L_{in}, which subsequently transmits them to a speech recognition model for audio-to-text encoding. Furthermore, a depth camera captures images of the workspace and the objects positioned on it, relaying them to a VLM. Utilizing provided captions, the VLM identifies and locates the objects on the board. Finally, another camera, directed towards the user, detects human body parts, denoted as P_{human}.

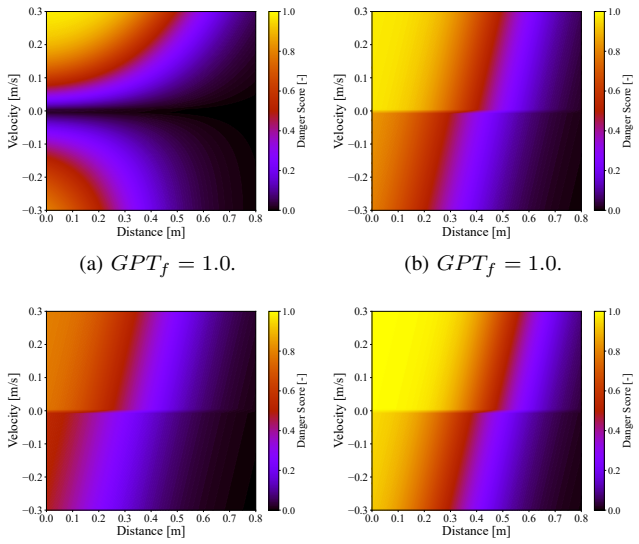
Safety evaluation: The safety evaluation proposed in this paper is based on an SI algorithm proposed by [18] by incorporating a scaling factor (GPT_f) representing the feeling of safety of the user. The scaling factor scales the danger score as shown in Equation 2. The baseline method [18] associates the danger to potential human injuries by calculating the impact severity of a collision and by anticipating the human motions to restrict the robot's speed closer to impact, which decreases the risk of a collision. The output is associated with the impact severity of the worst human action.

$$DS = \begin{cases} k_0 & \|\mathbf{d}\| = 0, \\ f_S(f_o \cdot GPT_f) & \|\mathbf{d}\| > 0, \frac{\pi}{2} < \phi \leq \pi, \\ f_S((1-\rho) f_o \cdot GPT_f)) & \|\mathbf{d}\| > 0, 0 \leq \phi \leq \frac{\pi}{2}, \end{cases} \tag{2}$$

where k_0 is a constant that is set to alert for potential collisions, f_S is a sigmoid function to ensure the Danger Score (DS) remains bounded between 0 and 1, f_o is a weighted sum of the speed of the end effector and the distance between the end effector of the robot and the human closest body part, GPT_f is a factor computed by an LLM to take into account the perception of safety, ρ is a human trust factor that ensures asymmetry in the DS function, ϕ is the angle between the direction vector and the relative velocity between the human and the robot end effector, $\|\mathbf{d}\|$ is the displacement vector between the end effector and the closest human body part, and $SI = \|DS - 1\|$. Fig. 3 shows how the proposed DS behaves with different GPT_f compared to the original DS.

The main difference with this method compared to the method proposed by [18] is that, in this proposed method, the DS adapts to different GPT_f; the danger remains high even at very low speeds (because of the imbalance in the weighted sum), and the danger decreases when the robot moves away from the robot.

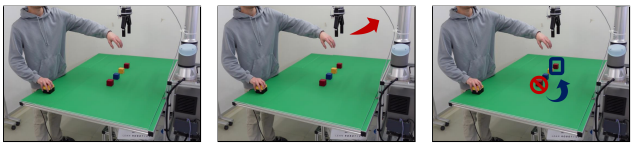
Decision-Making: The decision-making process takes place when the robot's speed is reduced to zero because the



(a) $GPT_f = 1.0$. (b) $GPT_f = 1.0$.

(c) $GPT_f = 0.5$. (d) $GPT_f = 1.5$.

Fig. 3. Relation of the Danger Score to the velocity and distance for a custom case. Velocity changes of sign when $\phi < \frac{\pi}{2}$. (a) the initial method proposed by [18]. (b) the proposed method with $GPT_f = 1$. On the bottom, (c) and (d), the proposed method with $GPT_f = 0.5$ and $GPT_f = 1.5$, respectively.



(a) Wait (h_1). (b) Withdraw (h_2). (c) Change Obj. (h_3).

Fig. 4. Decisions available for the robot: Wait (h_1), Withdraw (h_2), and Change Objective (h_3).

safety index is too low. Indeed, whenever the DS exceeds 0.8, indicating a significant level of risk, the robot is programmed to halt its movement. This ensures that the robot stops whenever the situation poses a high degree of danger, safeguarding against unwanted collisions. One way to decrease the DS is by asking the robot to move in the opposite direction of the human (moving in the opposite direction changes the angle ϕ). This characteristic of the SI can increase the system's efficiency and safety by preventing it from becoming stuck for an unnecessary amount of time. If a new objective in the opposite direction is suggested, the SI will decrease, and the robot will be able to move again. The user can at any time orally express his preferences regarding the decision the robot should make by modifying L_{in}. For example, the user might say "*When you are getting too close to me, always withdraw.*" As shown in Fig. 4, the robot has 3 functions he can choose from:

- h_1: the robot waits.
- h_2: the robot withdraws to an intermediate position.
- h_3: the robot changes the current objective.

Robot Engine: The LLM is used in two different ways to implement h_2. The first method consists of modifying the parking position suggested by [19]. This method uses a virtual force model to modify the end effector velocity and move it not only away from the human but also toward a parking position (P_{parking}), which can be asserted as a safer location. The LLM is used to suggest parking positions based on the user's input.

The second withdrawal method, unlike the first one, is not governed by any algorithm. The LLM directly suggests the withdrawal position in the Cartesian frame.

Similarly, h_3 is implemented with the LLM to decide which object the robot should pick up based on its current situation and the natural language prompt from the user. By directing the robot on which object to pick up, the LLM suggests a 3D pose that shapes a trajectory based on the user language input.

C. Post-Processing and Execution

LLMs, while powerful, produce stochastic answers and are prone to occasional hallucinations. In such scenarios, the constraint satisfaction module serves as a critical safeguard, ensuring that the robot operates within predefined constraints and boundaries despite the inherent uncertainties. This module rigorously adheres to manufacturer-defined constraints and safety design principles to guarantee the required level of safety at all times. The constraint satisfaction module ensures that the output position suggested by GPT-4 remains within the workspace and is not positioned in the direction of the user. Once the objective coordinates P_{obj} are processed by the constraint satisfaction module, the path of the robot is computed and executed.

IV. Experiments and Results

The first experiment investigates the integration of human safety perceptions into safety assessments by employing LLMs. The subsequent experiment delves into the accuracy of a streamlined coding paradigm, leveraging the generalization capabilities of GPT-4 to make decisions regarding the robot's subsequent actions based on the oral commands from the user. Finally, the last experiment seeks to validate the potential of LLMs in shaping trajectories by suggesting 3D poses influenced by multimodal input.

A. Experimental Setup

The setup involves a user and a UR5e robotic arm interacting across a table with 3D-printed cubes, which the robot manipulates using a two-finger Robotiq gripper. The system uses two Intel RealSense D435 cameras to monitor the scene. We implement the system using ROS Noetic on Ubuntu 20.04, and use MoveIt for motion planning. The developed software environment [20] integrates key open-source tools, including ROS and Docker, enabling automated testing and future work extensions.

The experiments use GPT-4 [21] for the interpretation and implementation of the user requests. The users' requests are translated from audio to text with Whisper [22]. For object and human body part detection, the system employs Detic [9] framework and BlazePose from MediaPipe [23], respectively.

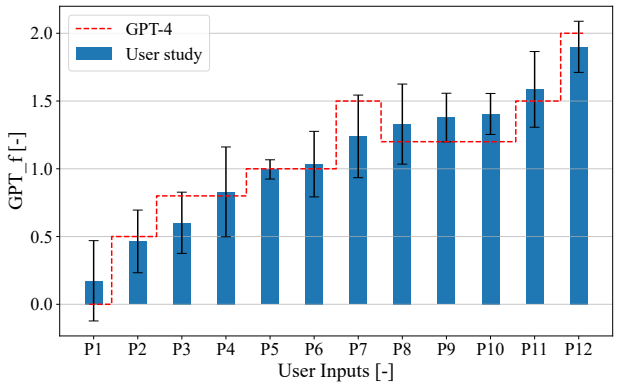


Fig. 5. Variation of GPT_f with different prompts: GPT_f in red is computed by GPT-4 and in blue, the mean computed through the user study with its standard deviation. The user inputs named P1 to P12 represent the 12 tested prompts.

The subjective evaluation was carried out via online forms we prepared. Depending on the experiment, we showed prompts, static diagrams, real-world videos, and/or animated plots. Then, we asked the participants to input their responses. All physical interactions between the human and the robot are demonstrated by the same person.

B. Safety Evaluation

This experiment is divided into a user study to compare user evaluation of GPT_f with GPT-4 estimation, and a quantitative assessment. 32 participants and GPT-4 are presented with the same initial prompt, both for elucidating the scenario and specifying the desired output format. They then receive the different natural language inputs, and the answers from the participants are collected through a designated form. Through a user study, the user is asked to suggest the GPT_f he/she deems most appropriate based on each prompt he/she is provided with. The user's mean and the answers from GPT-4 are used to compute the Root Mean Square Error (RMSE).

Firstly, as shown in Fig. 5, the output from GPT-4 has a similar increasing trend as the user. Another important point is that GPT-4 can emulate the understanding of the user's state of mind. For example, it suggests a higher factor when the user is feeling heartbroken (P11).

In the second part of the experiment, the behavior of the robot is analyzed with three different GPT_f across a spectrum of scenarios, ranging from easy to challenging. In the first scenario, cubes are symmetrically positioned on either side of the workspace, requiring the robot and user to retrieve objects from their respective sides. Both humans and robots are tasked with stacking the picked cubes on the same side as their retrieval. While this arrangement minimizes potential overlap between the human and robot workspaces, there exists a brief convergence when reaching for a centrally located cube. The second scenario increases complexity by centralizing all cubes on the table. In the third and most challenging scenario, the robot is tasked with retrieving and delivering cubes directly into the user's hand.

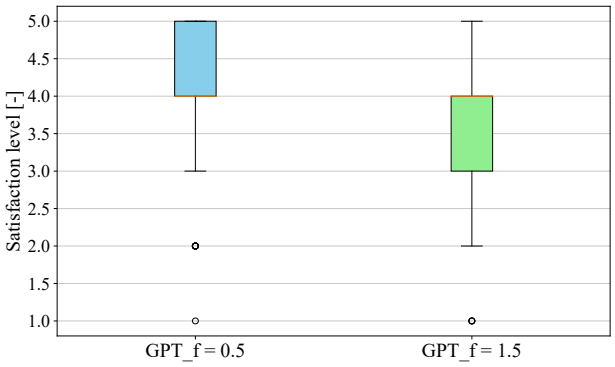


Fig. 6. Satisfaction analysis from the user study for the impact of using GPT_f = 0.5 and GPT_f = 1.5 on the robot behavior compared to a neutral behavior (GPT_f = 1.0). 1 is very unsatisfied, and 5 is very satisfied.

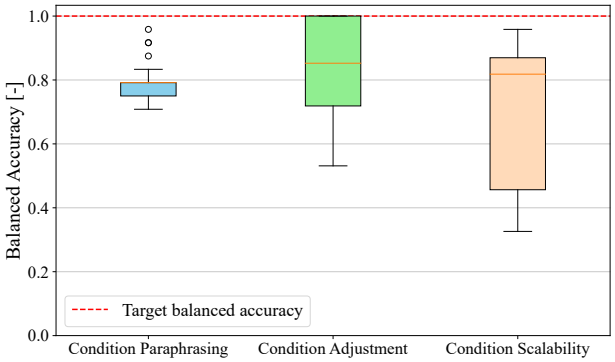


Fig. 7. Balanced accuracy of the model for the decision process based on natural language.

This configuration represents a pivotal aspect of collaborative interaction, simulating scenarios where proximity between the human and robot is imperative for task completion.

Fig. 6 shows that the satisfaction of both factors from the 32 users shares a similar median, but the satisfaction with $GPT_f = 1.5$ is slightly lower. Both distributions have a 75th percentile above 3, indicating that at least 75% of the responses for each factor are above the neutral satisfaction level of 3, showing generally positive assessments for both factors.

C. Decision-Making

The robot is halted in 24 different situations to assess the viability of utilizing LLMs for decision-making in HRC by providing a streamlined coding paradigm and evaluating its accuracy. These situations manipulate the input variables necessary for the decision-making process. Text constraints imposed on the input variables within the general prompt express the desired output, which is used to compute the decision-making accuracy by ensuring that the desired output aligns with the answer from GPT-4. Three parameters vary to create the different situations: the distance between the objective and the end effector of the robot, the number of cubes available to be picked up on the table, and the current action of the robot.

Three distinct tests are made to evaluate the viability of using LLMs as a streamlined coding paradigm. The first one consists of changing the natural language used to express the same condition on the input and thereby assessing the **Condition Paraphrasing**. For example, the initial condition is: "*Always change objective when picking and always wait when placing. Never withdraw.*" This condition is then modified to create new language inputs while preserving the original meaning. One of the alternative formulations is: "*Change the objective when picking and consistently wait when placing. Refrain from withdrawing.*" Then, different conditions are tested with language inputs that follow a similar structure, namely **Condition Adjustment**. The final step is to assess the scalability of the condition (**Condition Scalability**), gauging the system's capability to manage increasingly intricate conditions effectively.

As shown in Fig. 7, Condition Paraphrasing has the smallest standard deviation with a median slightly below the other medians. The Condition Adjustment displays a notably dispersed distribution, as shown by its higher standard deviation, with multiple balanced accuracy measurements nearing the 50% threshold. Condition scalability demonstrates elevated variance in comparison to Condition Adjustment, indicating that as the number of conditions increases, so does the variance.

D. Robot Engine

In addition to the two withdrawal methods detailed in Section III (Withdrawl 1 and Withdrawal 2), a third withdrawal method (Withdrawal 3) is used for the user study. The additional method is the original withdrawal method proposed by [19] with no impact from the user's oral commands. The three withdrawal methods are tested with a variety of prompts. These prompts express positions relative to the user's location or other objects on the board. The experiment is evaluated through a user base study of 30 participants who rate their satisfaction with each withdrawal method. For every prompt, they also need to justify what motivated their answer by selecting how significant the three factors were:

- Safety: *the robot should always move away from the user.*
- Compliance: *the robot should always follow the command from the user.*
- Human-like behavior: *the robot should exhibit behavior that closely resembles that of a human.*

Fig. 8 shows that Withdrawal 3 received the highest satisfaction ratings and the lowest dissatisfaction ratings. Withdrawal 1 and Withdrawal 2 show a smaller difference in their results. This is likely because Withdrawal 2 originates from Withdrawal 1, with only the change in parking position. Table I demonstrates that the differences between the distributions of the withdrawal methods are statistically significant with p-values below 0.01 and 0.001.

As shown in Fig. 9, according to the users' assessments, compliance was the most significant factor, followed by safety, which is expected because of the task's collaborative origin. Human-like behavior was generally deemed not significant by the users.

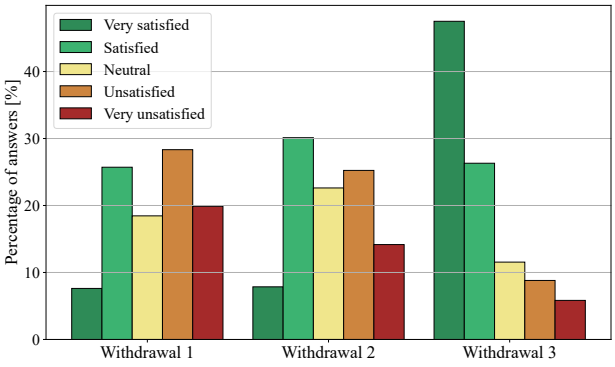


Fig. 8. Distributions of the answers from the user study for each withdrawal method. Withdrawal 1 is the original withdrawal method proposed by [19], Withdrawal 2 is the original withdrawal [19] method but augmented with the parking position modified by GPT-4, and Withdrawal 3 is the withdrawal method where GPT-4 directly suggests a withdrawal position.

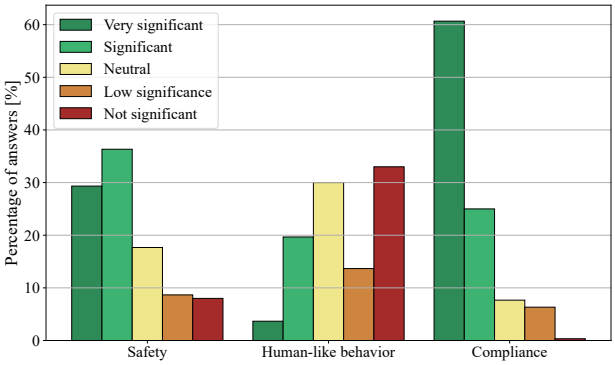


Fig. 9. Distributions of answers from the user study for each factor influencing the answers.

V. Discussion

The results show that GPT-4 can accurately predict user perceptions of safety, as demonstrated by the alignment between its predictions and user-reported states. Indeed, the robot adapted its behavior effectively based on the user's perception of safety, avoiding close proximity when users felt apprehensive, unlike when confidence around robots was detected. However, the number of prompts tested was limited to 12, and the indirect user-robot interaction from the user studies may have shaped the feedback from users in a unique way. Future experiments should include users with varying levels of comfort around robots to better validate these findings.

While GPT-4 performs well under simple conditions, its accuracy declines as conditions become more complex. As the balanced accuracy is not 100%, this system is more suitable for non-critical scenarios where errors have minimal consequences and should be used cautiously in high-stakes environments.

The robot accurately suggested 3D poses, aligning with

TABLE I

p-VALUES OF THE ANSWERS FROM THE USER STUDY FOR EACH WITHDRAWAL METHOD USING THE MANN-WHITNEY U TEST

	Withdrawal 1	Withdrawal 2	Withdrawal 3
Withdrawal 1	-	$*$	$**$
Withdrawal 2	$*$	-	$**$
Withdrawal 3	$**$	$**$	-

Significant differences of $p < 0.01$ are denoted by $*$ and $p < 0.001$ are denoted by $**$.

user choices in over 50% of trials. The method proved effective, though improvements are needed, particularly in withdrawal scenarios. Notably, users valued robot compliance over human-like behavior, emphasizing efficient collaboration. However, caution is warranted as the robot can sometimes generate erroneous poses. Non-generative algorithmic governance may strike a balance between safety and user needs, warranting further exploration.

VI. CONCLUSION

This work proposed a framework aimed at enhancing collaboration between humans and robots. Leveraging LLMs, the framework can realize adaptive safety evaluations based on users' safety perceptions, enabling the robot's behavior to adapt accordingly and ensure a safe collaboration. When faced with potentially hazardous situations, the robot employs a streamlined coding paradigm based on natural language input to autonomously determine its next course of action, thus facilitating seamless collaboration. Additionally, the proposed system shapes trajectories based on user preferences by suggesting 3D poses for the robot's end effector. We performed a subjective evaluation of the system with 32 subjects via online forms that show real-world videos, diagrams, prompts, and animated plots, as well as an objective evaluation for the accuracy of decision-making. The results show that GPTAlly infers safety perception similarly to humans, and achieves an average of 80% of accuracy in decision-making, with few instances under 50%.

As future work, the system could be enhanced by enabling it to learn from continuous user feedback, allowing it to adapt over time to evolving safety perceptions and collaboration preferences. Incorporating a reinforcement learning component would also enable the robot to refine both its decision-making processes and trajectory shaping based on real-time interactions and long-term user behavior trends. Moreover, a deeper exploration of the trade-offs between safety and productivity is essential. While the current framework focuses primarily on safety, its impact on task efficiency must be better understood, particularly in time-sensitive and high-efficiency environments.

REFERENCES

[1] E. Coronado, T. Kiyokawa, *et al.*, "Evaluating Quality in Human-Robot Interaction: A Systematic Search and Classification of Performance and Human-centered Factors, Measures and Metrics towards an Industry 5.0," *Journal of Manufacturing Systems*, vol. 63, pp. 392–410, Apr. 2022.

[2] C. D. Manning, "Human Language Understanding & Reasoning," *Daedalus: Journal of the American Academy of Arts and Sciences*, vol. 151, no. 2, pp. 127–138, May 2022.

[3] R. Firoozi, J. Tucker, *et al.*, "Foundation Models in Robotics: Applications, Challenges, and the Future," Dec. 2023. Preprint: https://doi.org/10.48550/arXiv.2312.07843

[4] J. Wang, E. Shi, *et al.*, "Prompt Engineering for Healthcare: Methodologies and Applications," Apr. 2023. Preprint: https://doi.org/10.48550/arXiv.2304.14670

[5] P. Georgiev, V. I. Lei, *et al.*, "Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context," Mar. 2024. Preprint: https://doi.org/10.48550/arXiv.2403.05530

[6] R. J. Kirschner, H. Mayer, *et al.*, "Expectable Motion Unit: Avoiding Hazards from Human Involuntary Motions in Human-Robot Interaction," *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 2, pp. 2993–3000, Apr. 2022.

[7] A. Zacharaki, I. Kostavelis, *et al.*, "Safety Bounds in Human Robot Interaction: A Survey," *Safety Science*, vol. 127, pp. 1–19 (104 667), July 2020.

[8] J. Liang, W. Huang, *et al.*, "Code as Policies: Language Model Programs for Embodied Control," Sept. 2022. Preprint: https://doi.org/10.48550/arXiv.2209.07753

[9] X. Zhou, R. Girdhar, *et al.*, "Detecting Twenty-Thousand Classes Using Image-Level Supervision," in *Proceedings of 17th European Conference on Computer Vision (ECCV 2022)*, S. Avidan, G. Brostow, *et al.*, Eds., Tel Aviv, Israel, Oct. 2022, pp. 350–368.

[10] Y. Chen, J. Arkin, *et al.*, "AutoTAMP: Autoregressive Task and Motion Planning with LLMs as Translators and Checkers," June 2023. Preprint: https://doi.org/10.48550/arXiv.2306.06531

[11] A. Bucker, L. Figueredo, *et al.*, "LATTE: LAnguage Trajectory TransformEr," in *Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA 2023)*, London, United Kingdom, May 2023, pp. 7287–7294.

[12] J. Devlin, M.-W. Chang, *et al.*, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018. Preprint: https://doi.org/10.48550/arXiv.1810.04805

[13] A. Radford, J. W. Kim, *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, vol. 139, (Virtual), July 2021, pp. 8748–8763.

[14] S. Haddadin, A. Albu-Schaffer, *et al.*, "The "DLR Crash Report": Towards a Standard Crash-Testing Protocol for Robot Safety - Part I: Results," in *Proceedings of 2009 IEEE International Conference on Robotics and Automation (ICRA 2009)*, Kobe, Japan, May 2009, pp. 272–279.

[15] S. Haddadin, A. Albu-Schaffer, *et al.*, "The "DLR Crash Report": Towards a Standard Crash-Testing Protocol for Robot Safety - Part II: Discussions," in *Proceedings of 2009 IEEE International Conference on Robotics and Automation (ICRA 2009)*, Kobe, Japan, May 2009, pp. 280–287.

[16] C.-S. Tsai, J.-S. Hu, *et al.*, "Ensuring Safety in Human-Robot Coexistence Environment," in *Proceedings of 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2014)*, Chicago, United States, Sept. 2014, pp. 4191–4196.

[17] K. Ikuta, H. Ishii, *et al.*, "Safety Evaluation Method of Design and Control for Human-Care Robots," *International Journal of Robotics Research (IJRR)*, vol. 22, no. 5, pp. 281–297, May 2003.

[18] G. A. Garcia Ricardez, A. Yamaguchi, *et al.*, "Human Safety Index based on Impact Severity and Human Behavior Estimation," in *Proceedings of 2nd International Conference on Mechatronics and Robotics Engineering (ICMRE 2016)*, Nice, France, Feb. 2016, pp. 177–190.

[19] G. A. Garcia Ricardez, A. Yamaguchi, *et al.*, "Withdrawal Strategy for Human Safety based on a Virtual Force Model," in *Proceedings of 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013)*, Tokyo, Japan, Nov. 2013, pp. 1119–1124.

[20] L. El Hafi, G. A. Garcia Ricardez, *et al.*, "Software Development Environment for Collaborative Research Workflow in Robotic System Integration," *RSJ Advanced Robotics (AR)*, vol. 36, no. 11, pp. 533–547, June 2022.

[21] J. Achiam, S. Adler, *et al.*, "GPT-4 Technical Report," Mar. 2023. Preprint: https://doi.org/10.48550/arXiv.2303.08774

[22] A. Radford, J. W. Kim, *et al.*, "Robust Speech Recognition via Large-Scale Weak Supervision," Dec. 2022. Preprint: https://doi.org/10.48550/arXiv.2212.04356

[23] S. Mroz, N. Baddour, *et al.*, "Comparing the Quality of Human Pose Estimation with BlazePose or OpenPose," in *Proceedings of 2021 International Conference on Bio-Engineering for Smart Technologies (BioSMART 2021)*, Paris, France, Dec. 2021, pp. 1–4.