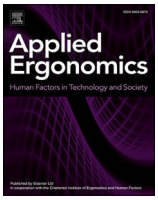
Applied Ergonomics 134 (2026) 104736

Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo



Exploring the integration of large language models in human-robot collaboration: Effects on performance, mental stress, and trust

Bingyi Su, Fangyuan Cheng, Lu Lu, Liwei Qing, SeHee Jung, Xu Xu *

Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, United States of America

ARTICLE INFO

Keywords:
Large language models
Human-robot collaboration
Performance
Mental stress
Trust

ABSTRACT

As large language models (LLMs) become increasingly integrated into robotic systems, understanding their influence on human-robot collaboration (HRC) is critical for designing effective and user-centered human-robot interactions. This study investigates the impact of LLM-enhanced robotic systems on users' performance, mental stress, and trust during collaborative tasks. Participants engaged in two representative HRC scenarios, including object delivery and instruction following, under two experimental conditions: with and without LLM support. Performance was measured through task completion time and number of verbal commands; mental stress was assessed using both subjective (NASA-TLX) and objective (galvanic skin response, GSR) measures; and trust was evaluated through the SHAPE Trust Index and eye-tracking metrics (blink rate and duration). Results showed that LLM integration significantly improved task efficiency and reduced subjective mental stress, particularly mental demand, effort, and frustration. Participants also reported higher levels of trust in the LLM condition across dimensions such as usefulness, reliability, accuracy, and ease of use. Interestingly, GSR data indicated elevated physiological arousal, possibly suggesting increased engagement or positive emotional activation, while eye-tracking measures showed no significant differences. These findings highlight the potential of LLMs to enhance HRC by enabling more natural communication, reducing mental workload, and increasing user trust, while also pointing to the need for improved system transparency to support deeper understanding and sustained trust.

1. Introduction

HRC has become an integral part of modern industrial and service applications, where robots are increasingly deployed to work alongside humans in diverse tasks. Future robotic systems are expected to transition from tools to teammates, operating and interacting naturally with humans in a manner resembling human–human teamwork (Ososky et al., 2013). Effective HRC relies on intuitive and efficient interaction methods. Current advancements in interaction modalities, such as gesture recognition and speech recognition, have played a significant role in improving the efficiency of HRC. However, such recognition systems often require users to memorize predefined commands, introducing additional cognitive and operational burdens. Studies have also shown that issuing verbal commands during HRC can be a source of stress; While verbal commands are generally preferred over gestures in assembly tasks, repeated verbal instructions can increase users' mental stress (Su et al., 2024). To promote seamless collaboration, it is crucial to explore innovative solutions that eliminate the need for memorizing specific commands.

LLMs have emerged as promising tools for addressing these challenges. Pretrained on extensive data, LLMs demonstrate exceptional natural language understanding and generation capabilities. Their ability to interpret user intentions and provide contextually appropriate responses has enabled their application across domains such as industry, education, healthcare, and entertainment. In the context of HRC, the integration of LLM models into robotics has emerged as a rapidly growing field, with researchers exploring their applications in perception, prediction, planning, and control (Firoozi et al., 2024). By leveraging LLMs, robots can better perceive and interact with the environment, facilitating the translation of high-level human instructions into actionable tasks. For example, Huang et al. (2022a) proposed a method to map natural language plans into executable actions, significantly improving task performance. Moreover, Singh et al. (2023) presented a LLM prompt structure for plan generation across environments, robot capabilities, and tasks. Similarly, Chen et al. (2024a) performed few-shot translation from natural language task

* Corresponding author.
E-mail address: xxu@ncsu.edu (X. Xu).

https://doi.org/10.1016/j.apergo.2026.104736
Received 22 April 2025; Received in revised form 12 January 2026; Accepted 15 January 2026
Available online 20 January 2026

descriptions to intermediate task representations to solve a robot task and motion plan using LLMs. In another study, researchers developed a framework to transform natural languages to temporal logics using LLMs (Chen et al., 2024b).

While these studies underscore the technical promise of LLMs in HRC, few have examined their impact on users' perceptions from a human factors perspective, especially related to mental stress and trust. Mental stress in the workplace is increasingly recognized as a critical factor that affects individual work performance and health (Umer, 2022). Therefore, integrating advanced technologies like LLMs into robotic systems necessitates a comprehensive evaluation of their effects on users' mental stress levels to ensure collaboration efficiency and worker well-being. Workers' mental stress is traditionally assessed through subjective measures, such as the NASA Task Load Index (NASA-TLX), which evaluates workload across cognitive and emotional dimensions (e.g., mental demand, effort, and frustration). However, self-reports are inherently limited by their post-hoc nature and reliance on self-awareness. To complement these limitations, objective physiological measures such as Galvanic Skin Response (GSR) offer continuous, real-time monitoring of stress-induced arousal by measuring skin conductance changes. The integration of both subjective and objective methods allows for a more comprehensive assessment of mental stress in HRC.

Equally important in HRC is trust, defined as the belief that an agent will act in ways that support the user's goals under uncertainty and vulnerability (Lee and See, 2004). In many cases, trust determines the extent to which humans accept and use robotic systems (Parasuraman and Riley, 1997). It influences system effectiveness, particularly as it is related to safety, performance, and use rate. Trust also influences a collaborator's willingness to assign tasks, share information, and cooperate, making it a pivotal factor in successful HRC (Freedy et al., 2007). Research suggests that higher levels of trust correlate with increased system usage, while insufficient trust often leads to disuse, hindering collaboration and performance (Khavas et al., 2020; Kok and Soh, 2020). As robots transition from tools to intelligent teammates, understanding trust in HRC is essential.

Subjective questionnaires remain the primary tools for evaluating trust in robotic systems. For instance, the SHAPE Automation Trust Index (Lewis et al., 2018) is a pragmatically oriented questionnaire that assesses trust across multiple dimensions, such as reliability, accuracy, and understanding. The Trust Perception Scale-HRI (Schaefer, 2016) was specifically designed to measure human-robot trust over time, expressed as an overall percentage of trust. Additionally, Charalambous et al. (2016) developed a psychometric scale to evaluate trust in industrial HRC, focusing on robot, human, and external aspects. Objective measures, as a complement, allow for real-time analysis of trust through physiological signals and behavioral data. Eye-tracking has been widely employed to evaluate trust in various contexts. For example, Jenkins and Jiang (2010) utilized eye-tracking data to evaluate operators' trust in human-robot interfaces in a context of urban search and rescue tasks. Similarly, Lu & Sarter (2019) investigated participants' trust levels using eye movement data in an unmanned aerial vehicle simulation. Other physiological measures, such as EEG, has been applied to model human trust in machines (Akash et al., 2018). Behavioral changes also serve as valuable indicators of trust. Studies by Lee et al. (2013) and Mota et al. (2016) demonstrated how behavioral observations can effectively measure trust in HRC.

The primary objective of this study is to investigate the impact of integrating LLMs into robotic systems on users' performance, mental stress, and trust during HRC. To this end, a robotic task planning system was developed using GPT-4o as the LLM-based AI assistant. The integrated LLM assistant interprets users' natural language verbal commands and generates corresponding motion plans for the robot to execute. The system was designed to support representative collaborative tasks with varying types and complexities, including object delivery and instruction following. Object delivery is one of the most fundamental and widely studied capabilities in HRC, which is often used as a canonical benchmark to evaluate coordination, timing, and shared intentions (Duan et al., 2024). This task requires the human-robot team to infer semantic meaning and understand high level task goals for collaboration, which typically imposes greater workloads (Liu et al., 2025; Tian et al., 2025). In contrast, instruction following tasks involve humans issuing sequential commands for the robot to execute. This paradigm places more emphasis on temporal planning and command decomposition but generally entails lower cognitive load and less semantic ambiguity (Cui et al., 2023; Shi et al., 2025). By including both tasks, this study enables examination of LLM integration effects across distinct interaction modalities and levels of task complexity. Through the combination of subjective assessments, physiological indicators, and behavioral data, this study aims to provide a comprehensive understanding of how LLM-enhanced systems influence human experience in collaborative robotic settings.

2. Methods

2.1. Participants

A total of 20 healthy participants (12 males, 8 females) aged 24–36 years (M = 27.7, SD = 2.74) were recruited from North Carolina State University. All participants were fluent English speakers with little to no prior experience in HRC. The experimental protocol was approved by the institutional review board at North Carolina State University (IRB #27542).

2.2. Experiment setup

In this experiment, collaborative tasks are carried out using a UR 5e robotic arm (Universal Robots), illustrated in Fig. 1.The robot's control and communication are managed through the ROS framework, operating within a Linux Ubuntu 20.04 environment.

Overall, this exploratory study employs a 2x2 factorial experimental design, where the independent variables are the presence of an LLM capability (two levels: with or without) and task type (two levels: object delivery and instruction following). In the collaborative object delivery scenario, participants worked alongside the robot to complete a circuit board assembly task. The robot was responsible for delivering items from a shelf containing 10 distinct objects necessary for various stages of assembly, including a circuit board, mounting frame, screwdriver, brush, and scissors. In conditions without LLM intelligence, participants were required to issue specific, predefined commands to request items. To enhance usability, the system allowed two fixed instruction formats: referring to the item by name (e.g., "deliver screwdriver") or by its assigned position number (e.g., "deliver item 3").In the LLM-enhanced condition, participants were allowed to use more natural and flexible language.For example, they could issue direct requests such as "please give me the screwdriver," or goal-oriented commands like "please give me the tool to tighten the screw," with the LLM assistant inferring the intended object (i.e., the screwdriver), locating it (e.g., position 3), and delivering it autonomously. The LLM also supported compound requests, enabling participants to issue combined statements such as "please pass me the screwdriver and the tool to cut the wire," which the assistant could interpret and execute sequentially by identifying and retrieving both the screwdriver and the scissors.

In the collaborative instruction-following scenario, participants directed the robot to place a Lego block into a designated blue box using verbal commands. This task required the execution of a sequence of actions: moving forward 20 cm, moving downward 20 cm, closing the gripper, moving left 50 cm, and opening the gripper. In the non-LLM condition, participants had to issue precise, step-by-step commands, e.g., "open gripper", with each command delivered individually and only after the robot had completed the preceding action. In contrast, the LLM-enhanced condition allowed participants to use natural language

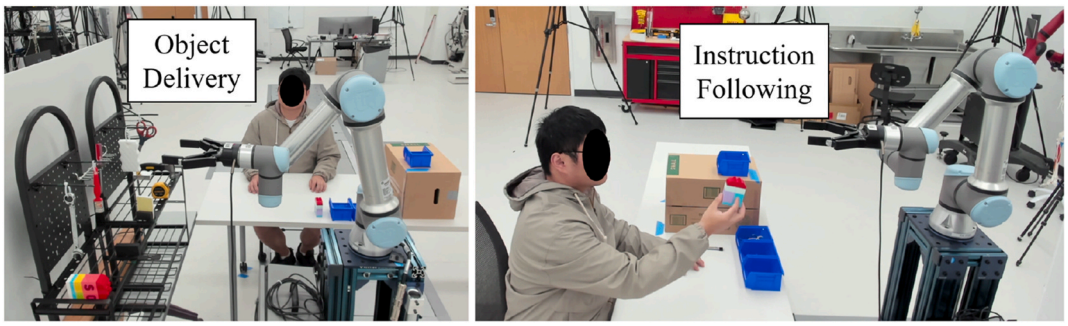


Fig. 1. Human-robot collaboration scenarios using a UR 5e robotic arm.

variations, such as "open your gripper" or "open your hand." Furthermore, participants could issue compound instructions, such as, "please move forward 20 cm, move down 20 cm, and then close your gripper", and the LLM assistant would automatically generate and execute a valid motion plan encompassing the full sequence. Using its embedded reasoning and generation capabilities, the LLM-based automatic mapping process ensured that natural-language instructions were consistently translated into executable actions, maintaining experimental reproducibility across participants.

2.3. LLM assistant design

The LLM-enhanced intelligence assistant in this study is built upon OpenAI GPT-4o, with the overall framework illustrated in Fig. 2. The workflow begins by converting the human operator's spoken natural language into textual input, enabling the AI assistant to process the instructions. The decision-making core of the assistant leverages OpenAI GPT-4o to interpret the input, evaluate contextual information, and address potential ambiguities. Based on its analysis, the assistant either generates conversational responses to request clarification from the human operator or produces a valid motion plan for the robotic arm to execute.

To optimize the assistant's reasoning capabilities and applicability across collaborative scenarios, several strategies were employed. A key enhancement involves the use of a few-shot prompt strategy, inspired by the structured hierarchical approach described by Ye et al. (2023). This strategy ensures a comprehensive understanding of both high-level task objectives and specific action details. The prompts are designed to progress from overall task descriptions to detailed robot function library information and operational constraints. By incorporating few-shot examples, the assistant's output is guided to align with the expected format and context for each scenario. An example of a few-shot prompt is provided in Fig. 3.

Moreover, in the collaborative object delivery scenario, the Retrieval Augmented Generation (RAG) strategy is implemented to identify and locate objects requested by the human operator. To support this process, a dedicated knowledge base is constructed, as shown in Fig. 4. This knowledge base includes detailed annotations for each object, encompassing attributes such as function, shape, color, material, purpose, and location. These annotations enhance the system's ability to accurately identify and retrieve the correct object based on human input.

2.4. Experiment procedure

The experiment is divided into three stages: pre-experiment, in-experiment, and post-experiment. In the pre-experiment stage, participants are provided with consent forms to review and sign. They are informed of their right to withdraw from the experiment at any time if they feel uncomfortable or unwilling to continue. Following this, the researcher explains the experiment in detail, including an introduction to the devices and tasks. After the introduction, a GSR sensor is attached to each participant to collect physiological data, and the eye-tracking headset is fitted and calibrated. Before the experiment, participants completed a structured training session to familiarize themselves with both the communication modes and the experimental tasks. In the non-LLM condition, participants were provided with a cheat sheet (Appendix 1, Section A, C) listing all predefined verbal commands and their corresponding robot actions. They were given sufficient time to review and memorize these commands to ensure accurate task execution. In the LLM-enabled condition, participants were instead shown several example interactions (Appendix 1, Section B, D) demonstrating how to flexibly instruct the robot using natural language. These examples

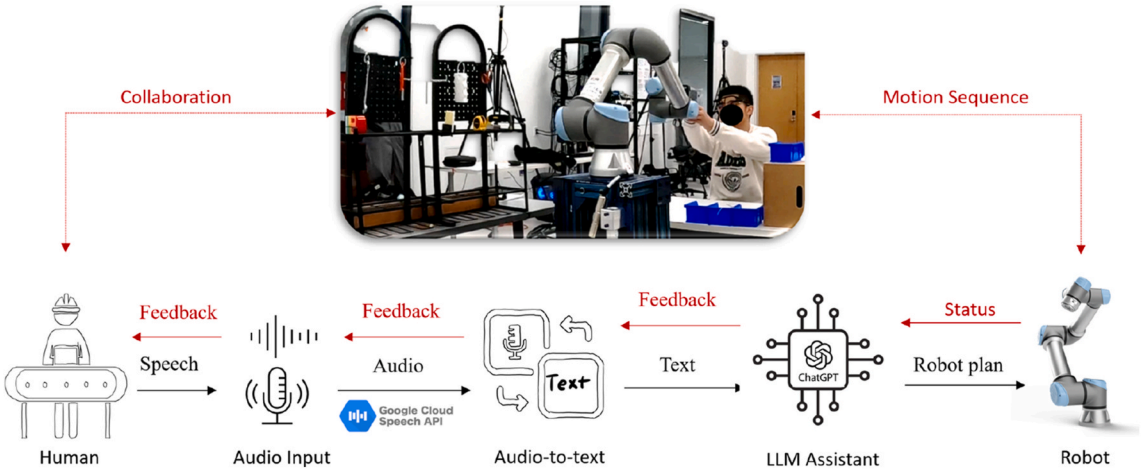


Fig. 2. The framework of the proposed LLM-enhanced robotic system.

Few-shot Prompt Strategy

Overview

You are my AI assistant for robotic task planning.

Task Description

We are collaborating to perform an assembly task involving object delivery using the robot's motion and control library.

Assistant Responsibilities

Your responsibility is to reason through my instructions, determine the required actions, and generate a sequence of valid motion commands for the robot arm.

RobotFuncLibrary

The robotic system you are working with is equipped with a comprehensive motion library that includes three types of functions: **robot arm motion**, **gripper motion**, and a **vision system**.
The detailed descriptions of these functions are as follows:
Robot Arm Motion Functions:

1. move_horizontally(d): This function moves the robotic arm laterally by a specified distance d, while maintaining the current vertical level of the arm.
2. ...

Constraints

Before generating the motion plan, it is essential to confirm that all the required information for each function is **available** and **valid**.
For example, you **must** determine the exact value of the distance d before using the move_horizontally(d) function. Additionally, **ensure** that the command parameters fall within the operational limits.
If any information is unclear or incomplete, proactively communicate with the human operator to obtain clarification.

Examples

User: Please move downwards 10cm, grasp the block, and deliver it to me.
Assistant:
- move_vertically(-0.1)
- close_gripper()
- move_to(person)
- open_gripper()

User: ...

Fig. 3. Example of the structured hierarchical few-shot prompt strategy.

Knowledge Base for Retrieval

Object 1: Scissors
Function: Cutting paper, fabric, or other materials.
Shape: Two elongated handles connected by a pivot, with two sharp, intersecting blades.
Color: Red handles with metallic silver blades.
Material: Plastic handles, stainless steel blades.
Purpose Annotation: Commonly used for office or crafting purposes.
Location: (x1,y1,z1)

Object 2: Screwdriver
Function: Tightening or loosening screws.
Shape: Cylindrical handle with a long, thin shaft ending in a flat head (flathead type).
Color: Blue handle with a black shaft.
Material: Rubber-coated handle for grip, steel shaft.
Purpose Annotation: Typically used for assembly tasks.
Location: (x2,y2,z2).
...

Fig. 4. Example of the knowledge base for retrieval.

demonstrated that the LLM assistant could interpret context-rich utterances, handle sequential instructions, and infer object identities based on semantic cues. Before data collection, the experimenter verified that each participant could correctly issue both predefined and free-form commands, and that the speech recognition system accurately transcribed all spoken inputs.

Participants are given a 10-min rest period before the start of the experiment to ensure they are in a relaxed state. Afterward, a 3-min baseline recording is collected to serve as a reference for physiological measurements. The in-experiment stage includes four sessions, representing a combination of two conditions. Latin Square design is used to mitigate time-order effects and reduce potential biases. Between each session, participants complete a self-report questionnaire and are given a 5-min rest period to recover.

In the post-experiment stage, wearable sensors are removed from the participants, and they are asked to sign a payment form to receive compensation for their participation. The total duration of the experiment is approximately 2 h.

2.5. Data collection and analysis

2.5.1. Performance measures

Completion time. Completion time refers to the total duration required for a participant to complete each task scenario, measured from the initiation of the first verbal command to the completion of the final robot action or task. This metric serves as a direct indicator of task efficiency and interaction fluency between the human and the robot (Steinfeld et al., 2006). To obtain accurate timing data, completion time was manually extracted from the video recordings of each experimental session.

Number of commands. The number of verbal commands issued by each participant during a task was used as a measure of communication efficiency (Hirsh et al., 2006). Fewer commands may reflect more effective and intuitive interaction. To quantify this metric, all commands and corresponding robot responses were automatically logged in a text file during the experiment, from which the total number of commands per task was extracted for analysis.

2.5.2. Mental stress measures

Subjective measure. NASA-TLX is a widely used subjective measure for mental stress. It performs a comprehensive assessment of mental workload using a multi-dimensional subscale, including mental demand, physical demand, temporal demand, performance, effort, and

frustration. Participants rate each dimension, and the overall mental workload is computed as the average of all subscales. This comprehensive approach captures various aspects of mental workload, providing insights into participants' subjective experiences of stress during the tasks.

Objective measure. Galvanic skin response is a well-established physiological indicator of mental stress. Unlike the skin conductance level, which represents a tonic measurement, the skin conductance response captures the phasic changes elicited by external stimuli (Dawson et al., 2007). Studies showed that palms, fingers, feet, and shoulders are effective locations for collecting the GSR data (van Dooren et al., 2012). As both hands are occupied in the experiment, the foot is chosen as the recording location using a Shimmer3 GSR + device.

The GSR device collects skin conductance data at a sampling rate of 256 Hz. The collected data is processed using MATLAB-based software, Ledalab (version 3.4.9). During processing, the signal is downsampled to 64 Hz and subjected to decomposition analysis to separate the phasic components from the tonic skin conductance levels (Benedek and Kaernbach, 2010). The mean values of the phasic components for each session are then used as indicators of mental stress levels. To account for individual physiological differences, the session values are normalized within each participant before statistical analysis. Specifically, each session's mean GSR value was divided by the participant's maximum mean value across all four sessions, as shown in Equation (1). This approach, consistent with prior research (Arai et al., 2010), standardizes inter-individual variability in physiological response magnitude.

$$GSR\_norm_{i,j} = \frac{GSR_{i,j}}{max_j\left(GSR_{i,j}\right)} \quad \text{(Eq. 1)}$$

where i represents the participant index (1–20) and j represents the experimental condition (1–4).

2.5.3. Trust measures

Subjective measure. The SHAPE Automation Trust Index (SATI) is a self-report questionnaire widely used to assess trust in automated systems, particularly in the fields of human-robot interaction and human factors research. SATI evaluates trust across six subscales: usefulness, reliability, accuracy, understanding, likability, and ease of use. Each subscale is rated on a scale from −5 to +5. The overall trust is computed as the average of all subscales.

Objective measure. Objective trust measurement involves analyzing raw eye movement data to extract metrics, including blink rate and blink duration. Objective trust measurement involves analyzing raw eye movement data to extract metrics, including blink rate and blink duration. These measures were selected as non-intrusive indicators of participants' trust during collaboration, as they are among the most widely used ocular indicators of attentional engagement and trust in HRC contexts. Prior research has shown that higher blink rate is associated with lower trust, while longer blink durations correspond to greater comfort and higher perceived trust (Ahmad and Alzahrani, 2023; Lu and Sarter, 2019; Ries et al., 2024; Zhang et al., 2024). Blink rate and duration were further chosen because they are robust to head motion, simple to extract, and suitable HRC tasks in which continuous gaze tracking is impractical. To capture these metrics, the Pupil Core system is utilized, comprising eye-tracking glasses equipped with high-speed cameras for monitoring eye movements and a scene camera to record the user's field of view. The system samples eye-tracking data at a frequency of 256 Hz.

Eye blinks are detected using the algorithm described by Hershman et al. (2018), which calculates noise measurements before and after missing-sample time windows to identify blink onset and offset. Similarly, to account for inter-individual variability, eye-blink metrics were normalized within each participant following the same procedure described in Equation (1).

2.5.4. Statistical analysis

A two-way repeated measures ANOVA will be performed in JMP to examine the effects of each factor on users' mental stress levels and trust levels, as well as their interaction effects. In addition, time order (session number: 1–4) was included as a covariate to account for potential residual time-order effects (e.g., practice or fatigue) not fully eliminated by the balanced Latin Square counterbalancing. Statistical significance will be determined at a threshold of 0.05.

3. Results

3.1. Performance measures

3.1.1. Completion time

There were significant main effects of both LLM capability ($F_{(1,19)} = 33.4978$, $p < .0001$) and task type ($F_{(1,19)} = 33.5804$, $p < .0001$), indicating that both factors independently influenced the time it took participants to complete the task. The interaction effect between LLM capability and task type was marginally non-significant ($F_{(1,19)} = 3.7541$, $p = .058$), suggesting a possible differential impact of LLM support depending on the task. Table 1 reports the four cell means for completion time across the LLM Presence × Task Type conditions; the fastest condition was instruction-following with LLM (91.75 ± 35.21 s), whereas the slowest condition was object delivery without LLM (254.20 ± 55.59 s). As illustrated in Fig. 5, participants completed tasks significantly faster when using LLM-enhanced systems (M = 145.98 ± 63.15 s) compared to traditional systems across all task types (M = 227.15 ± 87.72). Among task types, instruction-following tasks (M = 145.93 ± 95.15 s) exhibited a significantly shorter completion time than the object-delivery tasks (M = 227.20 ± 51.20 s).

3.1.2. Number of commands

The ANOVA results indicate a significant main effect of LLM capability on the number of verbal commands issued ($F_{(1,19)} = 112.97$, $p < .0001$), as well as a significant main effect of task type ($F_{(1,19)} = 24.33$, $p < .0001$). Table 2 reports the four cell means for the number of commands across the LLM Presence × Task Type conditions; the fewest commands were observed in the object-delivery task with LLM support (3.80 ± 0.83), whereas the most commands were observed in the instruction-following task without LLM support (15.00 ± 5.03). As illustrated in Fig. 6, participants issued significantly fewer commands in the LLM-enabled condition (M = 4.98 ± 1.99) than in non-LLM condition (M = 12.63 ± 5.00), and fewer commands were required for the object delivery task (M = 7.03 ± 4.23) than for the instruction following (M = 10.58 ± 5.88). However, the interaction between the two factors was not statistically significant ($F_{(1,19)} = 2.7796$, $p = .1013$).

3.2. Mental stress measures

3.2.1. NASA-TLX

There was a significant main effect of LLM capability on NASA-TLX scores ($F_{(1,19)} = 26.8656$, $p < .0001$). Table 3 reports the four cell means for NASA-TLX scores across the LLM Presence × Task Type conditions; the lowest workload was observed in the instruction-following task with LLM support (4.12 ± 2.98), whereas the highest workload was observed in the object-delivery task without LLM support (7.35 ± 2.98). As depicted in Fig. 7, participants reported consistently lower workload in the LLM-enabled condition (M = 4.28 ± 3.02) across all task types than non-LLM condition (M = 7.04 ± 3.48), indicating that

Table 1
Completion time by Condition LLM Presence × Task Type (M ± SD).

	Deliver	Follow
With LLM	200.20 ± 27.48	91.75 ± 35.21
Without LLM	254.20 ± 55.59	200.10 ± 105.67

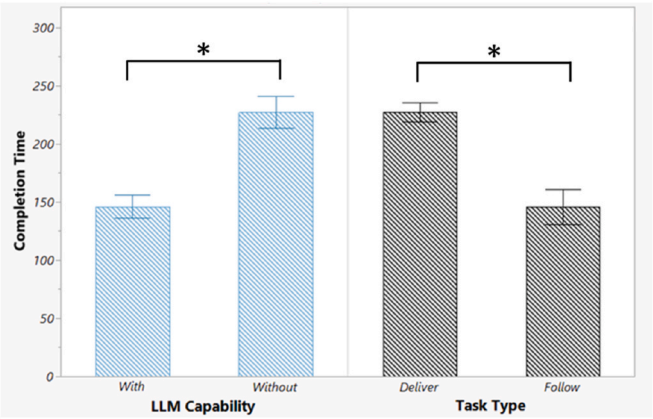


Fig. 5. Effects of LLM capability and task type on completion time.

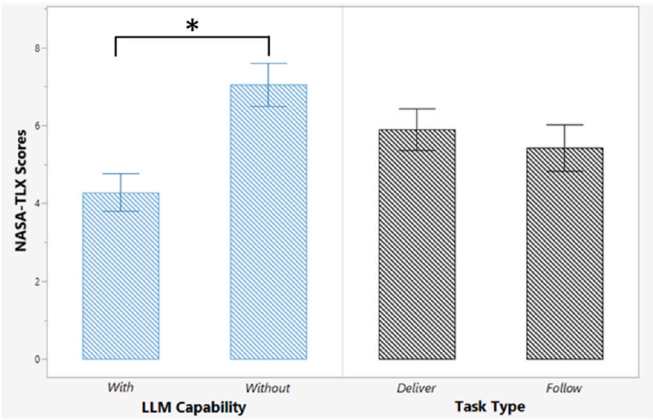


Fig. 7. Effects of LLM capability and task type on NASA-TLX scores.

Table 2
Number of commands by Condition LLM Presence × Task Type (M ± SD).

	Deliver	Follow
With LLM	3.80 ± 0.83	6.15 ± 2.13
Without LLM	10.25 ± 3.75	15.00 ± 5.03

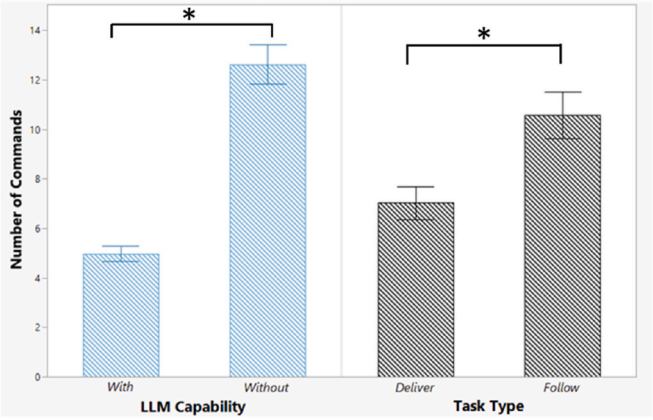


Fig. 6. Effects of LLM capability and task type on number of commands.

Table 3
NASA-TLX score by condition LLM presence × task type (M ± SD).

	Deliver	Follow
With LLM	4.45 ± 3.12	4.12 ± 2.98
Without LLM	7.35 ± 2.98	6.73 ± 3.98

the LLM-enhanced system reduced perceived cognitive demand. However, there are no significant effects of task type ($F_{(1,19)} = 0.7967$, p = .3760) and no significant effects for interaction ($F_{(1,19)} = 0.0709$, p = .7911).

To further examine the influence of LLM capability on subjective experiences of mental workload, ANOVA analysis was conducted for the six dimensions of the NASA-TLX. Significant main effects of LLM capability were found for mental demand, effort, performance, and frustration. Specifically, participants reported significantly lower mental demand in the LLM condition ($M = 5.93 \pm 4.90$) than in the non-LLM condition ($M = 9.70 \pm 5.03$), $F_{(1, 19)} = 26.71$, $p < .0001$. The required effort was also markedly lower with LLM support ($M = 3.75 \pm 3.11$) compared to without LLM assistance ($M = 7.45 \pm 4.31$), $F_{(1, 19)} = 25.73$, $p < .0001$. Participants further perceived their performance as improved under the LLM condition ($M = 2.93 \pm 2.96$) relative to the non-LLM condition ($M = 5.90 \pm 4.47$), $F_{(1, 19)} = 17.42$, $p = .0001$, and reported substantially reduced frustration ($M = 3.93 \pm 4.14$ vs. 8.20 ± 5.31), $F_{(1, 19)} = 38.02$, $p < .0001$. In contrast, no significant differences were observed for physical demand or temporal demand across conditions, suggesting that the impact of LLM integration was primarily confined to cognitive and emotional aspects of task execution rather than physical exertion or perceived time pressure. These results are depicted in Fig. 8.

3.2.2. GSR

The ANOVA results for normalized GSR values showed a significant main effect of LLM capability ($F_{(1,19)} = 19.7623$, p < .0001). Table 4 reports the four cell means for GSR across the LLM Presence × Task Type conditions; the highest GSR was observed in the instruction-following task with LLM support (0.795 ± 0.270), whereas the lowest GSR was observed in the instruction-following task without LLM support (0.460 ± 0.259). As seen in Fig. 9, participants exhibited significantly higher phasic GSR responses in LLM condition ($M = 0.79 \pm 0.29$) compared to the non-LLM condition ($M = 0.55 \pm 0.30$). Task type was not significant, with ($F_{(1,19)} = 2.0104$, p = .1620), and the interaction effect were not statistically significant, with ($F_{(1,19)} = 2.9352$, p = .092).

3.3. Trust measures

3.3.1. SHAPE Trust Index

There was a significant main effect of LLM capability on SHAPE Trust Index scores ($F_{(1,19)} = 43.4777$, p < .0001), indicating that participants trusted the LLM-enhanced system more. Table 5 reports the four cell means for SHAPE Trust Index scores across the LLM Presence × Task Type conditions; the highest trust ratings were observed in both LLM conditions (3.425 ± 0.878 for object delivery; 3.425 ± 0.979 for instruction following), whereas the lowest trust rating was observed in the instruction-following task without LLM support (1.858 ± 1.660). Fig. 10 highlights the elevated trust ratings in the LLM conditions (M = 3.43 ± 0.92 vs. 1.98 ± 1.43). Neither the main effect of task type, $F_{(1, 19)} = 0.30$, $p = .584$, nor the interaction between LLM capability and task type, $F_{(1, 19)} = 0.30$, $p = .584$, reached statistical significance.

To gain deeper insight into participants' trust perception, ANOVA analysis for the six dimensions of the SHAPE Trust Index was conducted. Significant main effects of LLM capability were found for usefulness ($F_{(1, 19)} = 30.03$, p < .0001; M = 3.53 ± 1.06 vs. 1.88 ± 2.08), reliability ($F_{(1, 19)} = 31.64$, p < .0001; M = 3.35 ± 1.05 vs. 1.83 ± 1.66), accuracy ($F_{(1, 19)} = 14.97$, p = .0003; M = 3.53 ± 1.24 vs. 2.25 ± 1.92), liking ($F_{(1, 19)} = 53.61$, p < .0001; M = 3.65 ± 1.19 vs. 1.15 ± 2.32), and ease of use ($F_{(1, 19)} = 25.20$, p < .0001; M = 3.43 ± 1.11 vs. 1.90 ± 2.19). These results indicate that LLM integration enhanced participants' perception of the robot's functionality, dependability, accuracy, likability, and usability. However, the understanding subscale did not show a statistically

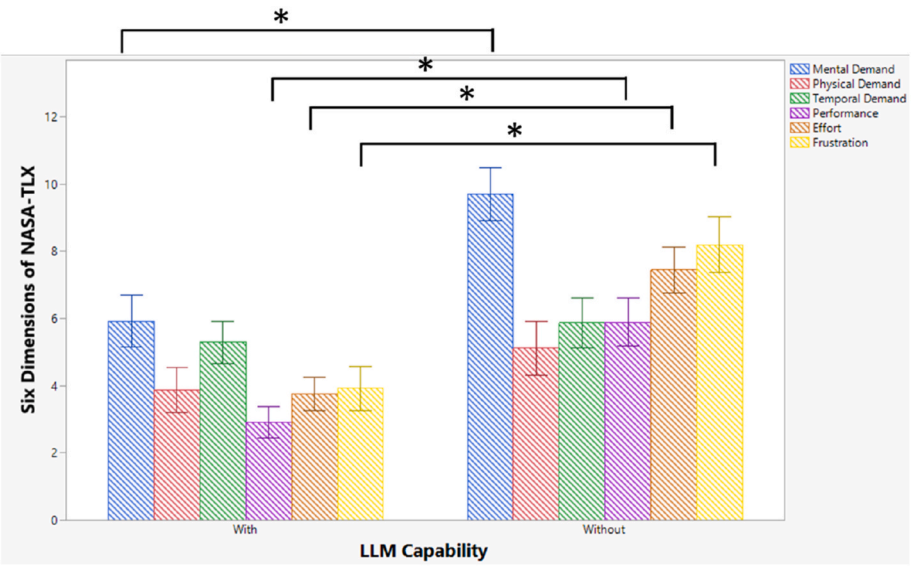


Fig. 8. Effects of LLM capability and task type on six dimensions of NASA-TLX scores.

Table 4
GSR by condition LLM presence × task type (M ± SD).

	Deliver	Follow
With LLM	0.779 ± 0.322	0.795 ± 0.270
Without LLM	0.630 ± 0.324	0.460 ± 0.259

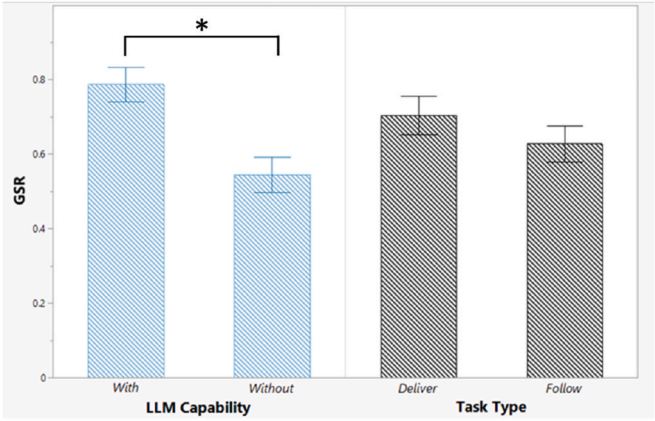


Fig. 9. Effects of LLM capability and task type on GSR

Table 5
SHAPE trust index by condition LLM presence × task type (M ± SD).

	Deliver	Follow
With LLM	3.425 ± 0.878	3.425 ± 0.979
Without LLM	2.10 ± 1.177	1.858 ± 1.660

significant difference between conditions, suggesting that the perceived interpretability or transparency of the robot's decision-making was not affected by LLM presence. These patterns are visually illustrated in Fig. 11.

3.4. Eye tracking

3.4.1. Blink rate

The ANOVA results indicate no significant main effects or interaction

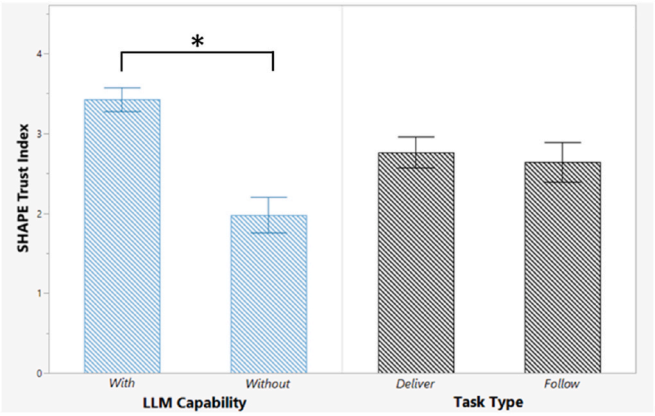


Fig. 10. Effects of LLM capability and task type on shape trust index scores.

for blink rate. The effect of LLM capability was not significant, $F_{(1, 19)} = 2.51$, $p = .122$, nor was the effect of task type, $F_{(1, 19)} = 1.43$, $p = .240$. The interaction between LLM capability and task type also failed to reach significance, $F_{(1, 19)} = 1.09$, $p = .304$. Table 6 reports the four cell means for blink rate across the LLM Presence × Task Type conditions; the lowest blink rate was observed in the object-delivery task with LLM support (0.668 ± 0.301), whereas the highest blink rate was observed in the instruction-following task without LLM support (0.847 ± 0.241). Fig. 12 shows relatively stable blink rates across all conditions, with a slight (but not statistically significant) trend toward lower blink frequency in the LLM condition.

3.4.2. Blink duration

Similarly, ANOVA results reveal no significant effects for blink duration across conditions. The effect of LLM capability was not significant, $F_{(1, 19)} = 0.49$, $p = .490$, nor was the effect of task type, $F_{(1, 19)} = 0.36$, $p = .554$. The interaction between LLM capability and task type was also non-significant, $F_{(1, 19)} = 0.11$, $p = .742$. Table 7 reports the four cell means for blink duration across the LLM Presence × Task Type conditions; the longest blink duration was observed in the object-delivery task with LLM support (0.793 ± 0.257), whereas the shortest blink duration was observed in the object-delivery task without LLM support (0.718 ± 0.334). Fig. 13 shows slight variations, but none were significant.

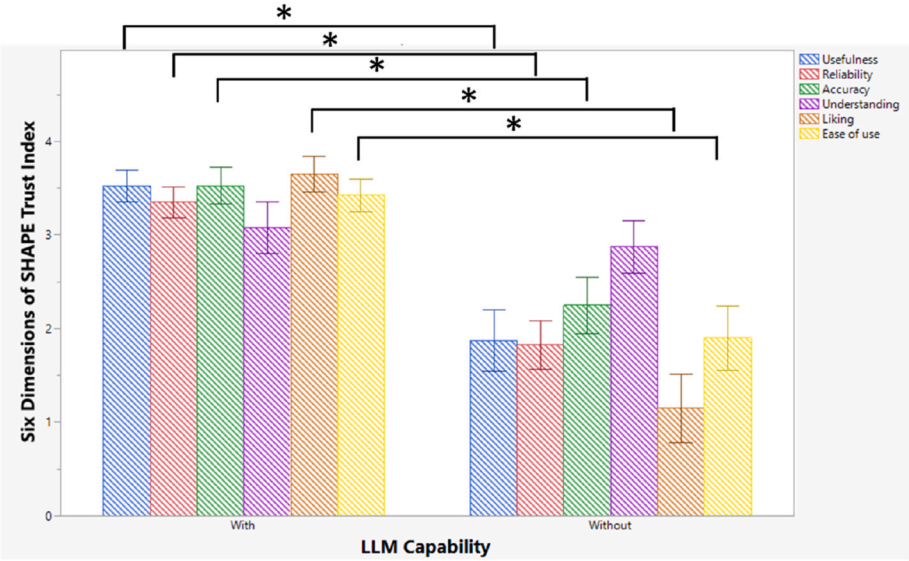


Fig. 11. Effects of LLM capability and task type on shape trust index scores.

Table 6
Blink rate by condition LLM presence × task type (M ± SD).

	Deliver	Follow
With LLM	0.668 ± 0.301	0.821 ± 0.263
Without LLM	0.833 ± 0.207	0.847 ± 0.241

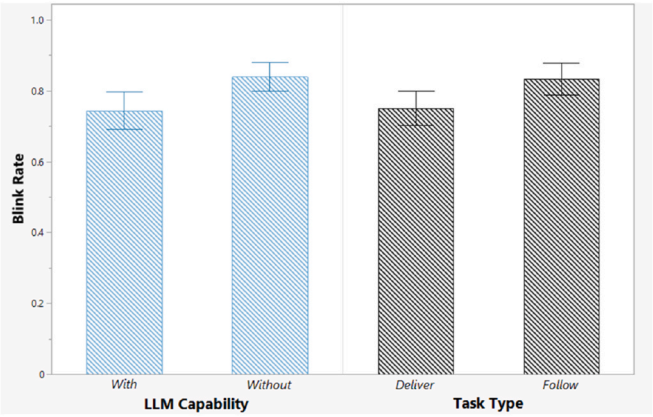


Fig. 12. Effects of LLM capability and task type on shape trust index scores.

Table 7
Blink duration by condition LLM presence × task type (M ± SD).

	Deliver	Follow
With LLM	0.793 ± 0.257	0.755 ± 0.308
Without LLM	0.718 ± 0.334	0.730 ± 0.334

4. Discussion

This study offers one of the earliest empirical examinations of LLM integration into human-robot collaboration from a human-factors perspective. Whereas previous work has largely centered on algorithmic performance or simulated interaction, we investigated how an LLM-driven robotic assistant affects users' task performance, mental workload, and trust during physical collaboration. The experimental framework uniquely combines subjective self-reports with objective

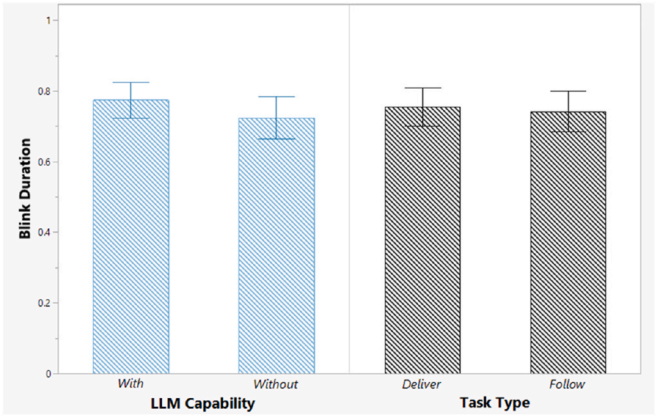


Fig. 13. Effects of LLM capability and task type on shape trust index scores.

physiological and behavioral indicators, yielding a holistic understanding of users' cognitive and affective states. The implementation of a real-time GPT-4o-based planning system employing few-shot prompting and retrieval-augmented grounding further distinguishes this work as a functional prototype of embodied, language-mediated collaboration. Beyond demonstrating measurable performance benefits, the findings highlight how LLM-driven natural language interaction can reduce cognitive effort and enhance user trust, offering new insights into the design of transparent and user-centered intelligent robotic systems.

The performance improvements, evidenced by significantly reduced task completion time and fewer commands issued in the LLM condition, align with findings from previous studies that emphasize the benefits of more intuitive interaction methods. For instance, Huang et al. (2022b) and Singh et al. (2023) demonstrated that LLMs can successfully translate natural language instructions into executable robot actions, streamlining task planning and execution. Similarly, Yao et al. (2024) showed that few-shot LLM prompting can improve the robot's ability to interpret complex user instructions in varied contexts. The present study extends these findings by providing empirical evidence from human-subject experiments, showing that such LLM-driven improvements translate into measurable performance gains in real-time HRC tasks.

In terms of mental stress, subjective ratings using the NASA-TLX

showed significant reductions under the LLM condition. Participants reported lower levels of mental demand, effort, and frustration, as well as improved self-assessed performance when interacting with the LLM-enhanced system. These findings is in line with a previous work (Su et al. (2024)) where it was found that although verbal commands are generally preferred over gestures, the need to remember and repeat rigid command structures can increase users' cognitive workload. The present study extends this insight by demonstrating that LLM integration supports more natural and flexible communication, thereby reducing the cognitive burden associated with issuing instructions.

However, a notable discrepancy emerged when comparing subjective stress ratings with objective physiological data. Contrary to the subjective reports, phasic GSR values were significantly higher in the LLM condition. This divergence suggests that while participants felt less mentally stressed, their physiological arousal increased during interaction with the LLM-enhanced system. One possible explanation is that this increase in arousal does not reflect negative stress, but rather heightened cognitive engagement or positive emotional resonance. As noted in previous research, GSR is a general indicator of arousal, sensitive not only to stress but also to positive affective states such as excitement, interest, and enjoyment (Boucsein, 2012). Supporting this interpretation, participants rated the LLM condition significantly lower on the "frustration" subscale of the NASA-TLX, with a mean score of 3.93 on a 0 to 20 scale. Moreover, the "liking" subscale of the SHAPE Trust Index received a significantly higher rating in the LLM condition, with a mean score of 3.65 on a scale ranging from −5 to 5. In this context, the elevated GSR values may reflect a more enjoyable and emotionally positive experience rather than a stressful one.

This divergence between perceived stress and physiological arousal underscores the complexity of interpreting multimodal data in HRC research. Subjective self-reports capture participants' reflective assessments, whereas physiological signals reveal underlying arousal levels that may arise from either positive or negative experiences. Taken together, these findings emphasize the importance of using both subjective and objective measures to gain a comprehensive understanding of user experience. Future research should consider complementing GSR data with additional indicators of emotional valence, such as facial expression analysis or heart rate variability, to better differentiate between stress and positive engagement in response to LLM-enabled robotic systems. Moreover, although subjective results indicated reduced mental workload, the relatively simple task structure may have limited physiological differentiation between conditions. Future studies should incorporate tasks with graded levels of difficulty, such as multi-step assembly, dynamic coordination, or safety-critical scenarios, to more comprehensively assess the influence of LLM assistance on performance, mental workload, and trust.

The significant increase in subjective trust observed in the LLM condition is consistent with findings from Akash et al. (2019), who reported that enhanced communication and system reliability are key contributors to increased trust in automated systems. In the present study, participants rated the LLM-enhanced system significantly higher across five dimensions of the SHAPE Trust Index, including usefulness, reliability, accuracy, liking, and ease of use. These results also align with the study by Ye et al. (2023), who reported that integrating ChatGPT into a virtual reality HRC environment significantly increased users' trust, primarily due to the system's ability to engage in more intuitive and effective communication.

Nevertheless, this study identified limitations of LLM-based systems in fostering a deeper understanding of the robot's behavior. The lack of significant difference in the "understanding" subscale of the SHAPE Trust Index suggests that while participants experienced functional and affective improvements, they did not perceive the LLM-assisted system as more transparent or interpretable. This raises concerns regarding the explainability of LLM-generated decisions in robotic contexts. As emphasized by Alonso and de la Puente (2018), system transparency and user understanding are critical components of sustained trust in human-machine interaction. Therefore, future work should focus on enhancing the interpretability of LLM outputs to further strengthen long-term trust in LLM-enabled robotic systems.

Additionally, objective trust measures captured through eye tracking, specifically blink rate and blink duration, did not exhibit significant differences across conditions. While previous studies reported blink-related metrics as non-intrusive indicators of trust (Lu and Sarter, 2019; Ries et al., 2024), with higher blink rates and longer blink durations often associated with decreased trust, the current study observed only similar directional trends without reaching statistical significance. One likely explanation is the relatively short duration of the tasks, which averaged around 3 min. This limited timeframe may not have been sufficient to elicit meaningful variation in ocular behavior to reveal trust-related differences. In contrast, studies such as Lu and Sarter's UAV monitoring experiments and Hergeth et al.'s automated driving research (Hergeth et al., 2016) involved session lengths of 30 min or more, during which significant blink-based trust differences emerged as participants had more time to calibrate their confidence in the system based on repeated reliability cues. We also acknowledge that the selected blink-based indicators, while well-established, may have limited sensitivity in capturing subtle trust dynamics within this interaction-oriented HRC setting. Although these measures provide practical and non-intrusive advantages, future studies should integrate additional ocular metrics, such as fixation entropy and pupil dilation, to enable a more sensitive evaluation of trust dynamics in LLM-assisted HRC.

Overall, the findings contribute to the emerging evidence that LLMs enhance HRC by promoting more natural interactions, reducing users' mental workload, and fostering greater trust. However, several limitations should be acknowledged. First, the study was conducted in a controlled laboratory environment with relatively short task durations, approximately 3 min per scenario on average. This limited exposure may not fully capture the evolving dynamics of long-term or real-world HRC. In particular, such short-term interactions constrain the ability to observe phenomena like trust calibration (Jelínek and Fischer, 2023) and trust decay (Centeio Jorge et al., 2023). Future research should therefore involve longer-term deployments, ideally spanning multiple days or weeks in authentic industrial contexts, to assess how LLM-based assistance influences sustained trust over time. Second, the study focused on two relatively simple collaborative scenarios using a single UR 5e robot arm. Although significant differences were observed in task completion time and the number of user commands, there were no significant effects of task type on mental stress or trust metrics. More complex and safety-critical collaborations, such as multi-stage assembly or shared-workspace navigation involving mobile robots, are likely to introduce additional cognitive and physical demands (Nahid et al., 2024; Steinfeld et al., 2006). Replicating the study using a broader range of robots and task scenarios with increased complexity will be beneficial to generalize the findings and understand the boundaries of LLM-enhanced collaboration. Third, while participants reported higher affective trust, they did not perceive a clearer understanding of the robot's reasoning. This indicates a persistent lack of system transparency. Integrating explainable-AI techniques, such as step-wise plan visualizations or natural-language explanations, could improve users' comprehension of the robot's decision-making processes (Biessmann and Treu, 2021; Matarese et al., 2021). Future work should explore the integration of such features and empirically evaluate their impact on users' understanding and trust. Addressing these limitations is crucial for advancing the development and deployment of LLM-enhanced robotic systems in real-world applications.

5. Conclusion

This study examined the effects of integrating LLMs into robotic systems on users' performance, mental stress, and trust during HRC. Results showed that LLM-enhanced systems significantly improved performance by reducing task completion time and the number of

commands required, while reducing operators' perceived mental stress. Additionally, trust was notably higher in the LLM condition across multiple dimensions of the SHAPE Trust Index. Overall, these findings highlight the potential of LLMs to improve the efficiency, comfort, and trustworthiness of human-robot interaction, while also emphasizing the need for greater transparency and explainability in future systems.

CRediT authorship contribution statement

Bingyi Su: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Data curation, Conceptualization. **Fangyuan Cheng:** Writing – original draft, Validation, Investigation, Data curation, Conceptualization. **Lu Lu:** Methodology, Investigation, Conceptualization. **Liwei Qing:** Investigation, Conceptualization. **SeHee Jung:** Validation, Investigation. **Xu Xu:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This manuscript is based on work supported by the National Science Foundation under Grant #2024688.

Appendix 1. Command Sets and Examples Used for Each Experimental Condition

A. Instruction Following Commands (Without LLM Assistance)

Actions	Commands	Notes
Move left	*"move left"*	Move left 10 cm
Move right	*"move right"*	Move right 10 cm
Move up	*"move up"*	Move up 10 cm
Move down	*"move down"*	Move down 10 cm
Move forward	*"move forward"*	Move forward 10 cm
Move backward	*"move backward"*	Move backward 10 cm
Open gripper	*"open gripper"*	Open gripper
Close gripper	*"close gripper"*	Close gripper

B. Instruction Following Commands Examples (With LLM Assistance):

"Please move left"
"Please move right 30 cm"
"Open your hand, please"
"Close your gripper and then move to the left 40 cm"

C. Object Delivery Commands (without LLM Assistance)

Actions	Command 1	Command 2
Deliver the wrench	*"deliver wrench"*	"deliver item one"
Deliver the brush	*"deliver brush"*	"deliver item two"
Deliver the screwdriver	*"deliver screwdriver"*	"deliver item three"
Deliver the sponge	*"deliver sponge"*	"deliver item four"
Deliver the scissors	*"deliver scissors"*	"deliver item five"
Deliver the Lego block	*"deliver lego block"*	"deliver item six"
Deliver the mounting frame	*"deliver mounting frame"*	"deliver item seven"
Deliver the circuit board	*"deliver circuit board"*	"deliver item eight"
Deliver the tape measure	*"deliver tape measure"*	"deliver item nine"
Deliver the hammer	*"deliver hammer"*	"deliver item ten"

D. Object Delivery Commands Examples (with LLM Assistance):

"Give me the screwdriver and then deliver the mounting frame to me"
"Could you please pass me the circuit board?"
"Please give me the tool to cut the cable"
"I need to brush to clean the circuit board"

References

Ahmad, M., Alzahrani, A., 2023. Crucial clues: investigating psychophysiological behaviors for measuring trust in human-robot interaction. INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION 135–143. https://doi.org/10.1145/3577190.3614148.

Akash, K., Hu, W.-L., Jain, N., Reid, T., 2018. A classification model for sensing human trust in machines using EEG and GSR. ACM Trans. Interact. Intell. Syst. 8 (4), 27: 1–27:20. https://doi.org/10.1145/3132743.

Akash, K., Polson, K., Reid, T., Jain, N., 2019. Improving human-machine collaboration through transparency-based feedback – part I: human trust and workload model. IFAC-PapersOnLine 51 (34), 315–321. https://doi.org/10.1016/j.ifacol.2019.01.028.

Alonso, V., de la Puente, P., 2018. System transparency in shared autonomy: a mini review. Front. Neurorob. 12. https://doi.org/10.3389/fnbot.2018.00083.

Arai, T., Kato, R., Fujita, M., 2010. Assessment of operator stress induced by robot collaboration in assembly. CIRP Ann. 59 (1), 5–8. https://doi.org/10.1016/j.cirp.2010.03.043.

Benedek, M., Kaernbach, C., 2010. A continuous measure of phasic electrodermal activity. J. Neurosci. Methods 190 (1), 80–91. https://doi.org/10.1016/j.jneumeth.2010.04.028.

Biessmann, F., Treu, V., 2021. *A turing Test for transparency* (No. arXiv:2106.11394). arXiv. https://doi.org/10.48550/arXiv.2106.11394.

Boucsein, W., 2012. Electrodermal Activity. Springer US. https://doi.org/10.1007/978-1-4614-1126-0.

Centeio Jorge, C., Bouman, N.H., Jonker, C.M., Tielman, M.L., 2023. Exploring the effect of automation failure on the human's trustworthiness in human-agent teamwork. Frontiers in Robotics and AI 10. https://doi.org/10.3389/frobt.2023.1143723.

Charalambous, G., Fletcher, S., Webb, P., 2016. The development of a Scale to evaluate trust in industrial human-robot Collaboration. International Journal of Social Robotics 8 (2), 193–209. https://doi.org/10.1007/s12369-015-0333-8.

Chen, Y., Arkin, J., Dawson, C., Zhang, Y., Roy, N., Fan, C., 2024a. *AutoTAMP: autoregressive task and motion planning with LLMs as translators and checkers* (No. arXiv:2306.06531). arXiv. http://arxiv.org/abs/2306.06531.

Chen, Y., Gandhi, R., Zhang, Y., Fan, C., 2024b. *NL2TL: transforming natural languages to temporal logics using large Language models* (No. arXiv:2305.07766). arXiv. http://arxiv.org/abs/2305.07766.

Cui, Y., Karamcheti, S., Palleti, R., Shivakumar, N., Liang, P., Sadigh, D., 2023. "No, to the Right"—Online Language corrections for robotic manipulation via shared autonomy. Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction 93–101. https://doi.org/10.1145/3568162.3578623.

Dawson, M.E., Schell, A.M., Filion, D.L., 2007. The electrodermal system. In: Handbook of Psychophysiology, third ed. Cambridge University Press, pp. 159–181. https://doi.org/10.1017/CBO9780511546396.007.

Duan, H., Yang, Y., Li, D., Wang, P., 2024. Human–robot object handover: recent progress and future direction. Biomimetic Intelligence and Robotics 4 (1), 100145. https://doi.org/10.1016/j.birob.2024.100145.

Firoozi, R., Tucker, J., Tian, S., Majumdar, A., Sun, J., Liu, W., Zhu, Y., Song, S., Kapoor, A., Hausman, K., Ichter, B., Driess, D., Wu, J., Lu, C., Schwager, M., 2024. Foundation models in robotics: applications, challenges, and the future. Int. J. Robot Res., 02783649241281508 https://doi.org/10.1177/02783649241281508.

Freedy, A., DeVisser, E., Weltman, G., Coeyman, N., 2007. Measurement of trust in human-robot collaboration. 2007 International Symposium on Collaborative Technologies and Systems, pp. 106–114. https://doi.org/10.1109/CTS.2007.4621745.

Hergeth, S., Lorenz, L., Vilimek, R., Krems, J.F., 2016. Keep your scanners peeled: gaze behavior as a measure of automation trust during highly automated driving. Hum. Factors 58 (3), 509–519. https://doi.org/10.1177/0018720815625744.

Hershman, R., Henik, A., Cohen, N., 2018. A novel blink detection method based on pupillometry noise. Behav. Res. Methods 50 (1), 107–114. https://doi.org/10.3758/s13428-017-1008-1.

Hirsh, R., Graham, J., Tyree, K.S., Sierhuis, M., Clancey, W., 2006. Intelligence for human-assistant planetary surface robots. https://www.semanticscholar.org/paper/Intelligence-for-Human-Assistant-Planetary-Surface-Hirsh-Graham/4046e60bf173dddb4d9a076cea829f885f7f241c.

Huang, W., Abbeel, P., Pathak, D., Mordatch, I., 2022a. *Language models as zero-shot planners: extracting actionable knowledge for embodied agents* (No. arXiv:2201.07207). arXiv. http://arxiv.org/abs/2201.07207.

Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., Sermanet, P., Brown, N., Jackson, T., Luu, L., Levine, S., Hausman, K., Ichter, B., 2022b. Inner monologue: embodied reasoning through planning with Language models. No. arXiv:2207.05608). https://doi.org/10.48550/arXiv.2207.05608 arXiv.

Jelínek, M., Fischer, K., 2023. *Trust regulation in social robotics: from violation to repair* (No. arXiv:2304.10360). arXiv. https://doi.org/10.48550/arXiv.2304.10360.

Jenkins, Q., Jiang, X., 2010. Measuring Trust and Application of Eye Tracking in Human Robotic Interaction.

Khavas, Z.R., Ahmadzadeh, S.R., Robinette, P., 2020. Modeling trust in Human-Robot Interaction: a Survey. In: Wagner, A.R., Feil-Seifer, D., Haring, K.S., Rossi, S., Williams, T., He, H., Sam Ge, S. (Eds.), Social Robotics. Springer International Publishing, pp. 529–541. https://doi.org/10.1007/978-3-030-62056-1_44.

Lee, J.D., See, K.A., 2004. Trust in automation: designing for appropriate reliance. Hum. Factors 46 (1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392.

Kok, B.C., Soh, H., 2020. Trust in robots: Challenges and opportunities. Curr. Robot. Rep. 1 (4), 297–309. https://doi.org/10.1007/s43154-020-00029-y.

Lee, J.J., Knox, B., Baumann, J., Breazeal, C., DeSteno, D., 2013. Computationally modeling interpersonal trust. Front. Psychol. 4. https://doi.org/10.3389/fpsyg.2013.00893.

Lewis, M., Sycara, K., Walker, P., 2018. The role of trust in human-robot interaction. In: Abbass, H.A., Scholz, J., Reid, D.J. (Eds.), Foundations of Trusted Autonomy, 117. Springer International Publishing, pp. 135–159. https://doi.org/10.1007/978-3-319-64816-3_8.

Liu, C., Wang, W., Li, C., Pang, R., Shang, Y., Gao, Y., 2025. Human-to-robot handovers based on multimodal perception. Neurocomputing 645, 130435. https://doi.org/10.1016/j.neucom.2025.130435.

Lu, Y., Sarter, N., 2019. Eye tracking: a process-oriented method for inferring trust in automation as a function of priming and System reliability. IEEE Transactions on Human-Machine Systems 49 (6), 560–568. https://doi.org/10.1109/THMS.2019.2930980. IEEE Transactions on Human-Machine Systems.

Matarese, M., Rea, F., Sciutti, A., 2021. A user-centred framework for explainable artificial intelligence in human-robot interaction. arXiv:2109.12912). https://doi.org/10.48550/arXiv.2109.12912 arXiv.

Mota, R.C.R., Rea, D.J., Le Tran, A., Young, J.E., Sharlin, E., Sousa, M.C., 2016. Playing the 'trust game' with robots: social strategies and experiences. 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 519–524. https://doi.org/10.1109/ROMAN.2016.7745167.

Nahid, N., Xinyi, M., Inoue, S., Ahad, M.A.R., 2024. Psychological analysis in human-robot Collaboration from workplace stress factors: a review, 165–195. https://doi.org/10.1201/9781032636054-11.

Ososky, S., Schuster, D., Phillips, E., Jentsch, F., 2013. Building Appropriate Trust in Human-Robot Teams.

Parasuraman, R., Riley, V., 1997. Humans and automation: use, misuse, disuse, abuse. Hum. Factors 39 (2), 230–253. https://doi.org/10.1518/001872097778543886.

Ries, A.J., Aroca-Ouellette, S., Roncone, A., de Visser, E.J., 2024. Gaze-Informed Signatures of Trust and Collaboration in Human-Autonomy Teams.

Schaefer, K.E., 2016. Measuring trust in human robot interactions: development of the "Trust Perception Scale-HRI.". In: Mittu, R., Sofge, D., Wagner, A., Lawless, W.F. (Eds.), Robust Intelligence and Trust in Autonomous Systems. Springer US, pp. 191–218. https://doi.org/10.1007/978-1-4899-7668-0_10.

Shi, L.X., Ichter, B., Equi, M., Ke, L., Pertsch, K., Vuong, Q., Tanner, J., Walling, A., Wang, H., Fusai, N., Li-Bell, A., Driess, D., Groom, L., Levine, S., Finn, C., 2025. Hi robot: Open-Ended instruction following with hierarchical vision-language-action models. No. arXiv:2502.19417). https://doi.org/10.48550/arXiv.2502.19417 arXiv.

Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., Garg, A., 2023. ProgPrompt: generating situated robot task plans using large Language models. 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 11523–11530. https://doi.org/10.1109/ICRA48891.2023.10161317.

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., Goodrich, M., 2006. Common metrics for human-robot interaction. Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction, pp. 33–40. https://doi.org/10.1145/1121241.1121249.

Su, B., Jung, S., Lu, L., Wang, H., Qing, L., Xu, X., 2024. Exploring the impact of human-robot interaction on workers' mental stress in collaborative assembly tasks. Appl. Ergon. 116, 104224. https://doi.org/10.1016/j.apergo.2024.104224.

Tian, L., Xu, S., He, K., Love, R., Cosgun, A., Kulic, D., 2025. *Collaborative object handover in a robot crafting assistant* (No. arXiv:2502.19991). arXiv. https://doi.org/10.48550/arXiv.2502.19991.

Umer, W., 2022. Simultaneous monitoring of physical and mental stress for construction tasks using physiological measures. J. Build. Eng. 46, 103777. https://doi.org/10.1016/j.jobe.2021.103777.

van Dooren, M., de Vries, J.J.G., Gert-J, Janssen, J.H., 2012. Emotional sweating across the body: comparing 16 different skin conductance measurement locations. Physiol. Behav. 106 (2), 298–304. https://doi.org/10.1016/j.physbeh.2012.01.020.

Yao, B., Chen, G., Zou, R., Lu, Y., Li, J., Zhang, S., Sang, Y., Liu, S., Hendler, J., Wang, D., 2024. More samples or more prompts? Exploring effective few-shot In-Context learning for LLMs with In-Context sampling. Findings of the Association for Computational Linguistics 1772–1790. https://doi.org/10.18653/v1/2024.findings-naacl.115. *NAACL 2024*.

Ye, Y., You, H., Du, J., 2023. Improved trust in human-robot Collaboration with ChatGPT. IEEE Access 11, 55748–55754. https://doi.org/10.1109/ACCESS.2023.3282111. IEEE Access.

Zhang, Y., Yadav, A., Hopko, S.K., Mehta, R.K., 2024. Gaze we trust: comparing eye tracking, self-report, and physiological indicators of dynamic trust during HRI. Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, pp. 1188–1193. https://doi.org/10.1145/3610978.3640649.