How Do We Research Human-Robot Interaction in the Age of Large Language Models? A Systematic Review

Yufeng Wang*
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
Zhejiang University
Hangzhou, China
wyufeng@zju.edu.cn

Yuan Xu*
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
yxu712@connect.hkust-gz.edu.cn

Anastasia Nikolova
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
anikolova721@connect.hkust-gz.edu.cn

Yuxuan Wang
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
Savannah College of Art and Design
Savannah, USA
yuwang85@student.scad.edu

Jianyu Wang
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
Zhejiang University
Hangzhou, China
3220100890@zju.edu.cn

Chongyang Wang†
West China Hospital
Sichuan University
Chengdu, China
mvrjustid@gmail.com

Xin Tong†
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
The Hong Kong University of Science and Technology
Hong Kong, China
xint@hkust-gz.edu.cn

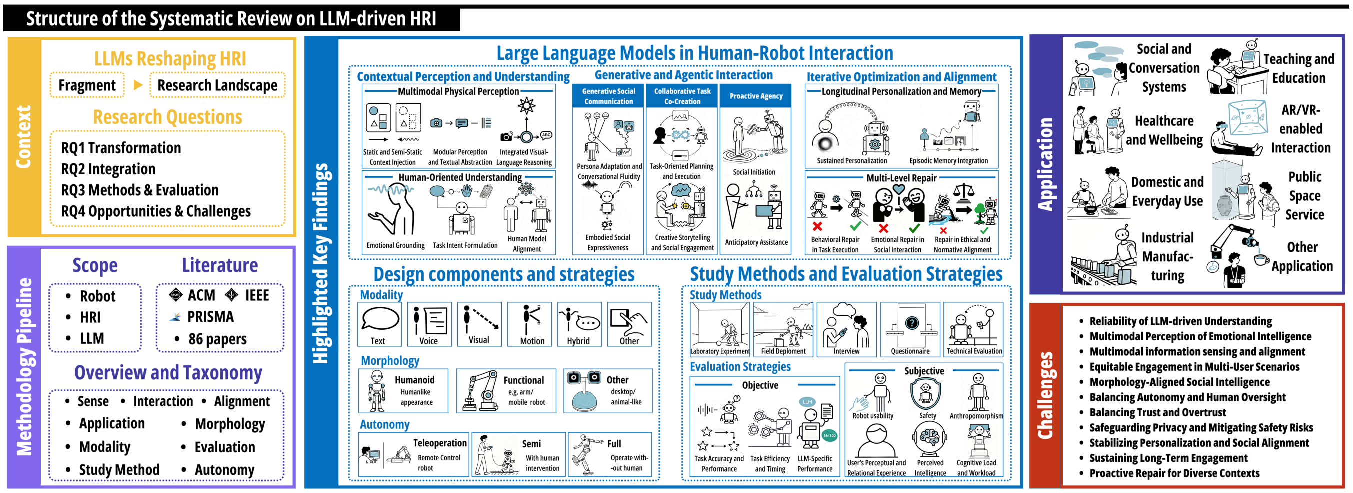

Structure of the Systematic Review on LLM-driven HRI
Context
LLMs Reshaping HRI
Fragment
Research Landscape
Research Questions
RQ1 Transformation
RQ2 Integration
RQ3 Methods & Evaluation
RQ4 Opportunities & Challenges
Methodology Pipeline
Scope
• Robot
• HRI
• LLM
Literature
ACM
IEEE
PRISMA
• 86 papers
Overview and Taxonomy
• Sense
• Interaction
• Alignment
• Application
• Morphology
• Modality
• Evaluation
• Study Method
• Autonomy
Highlighted Key Findings
Large Language Models in Human-Robot Interaction
Contextual Perception and Understanding
Multimodal Physical Perception
Static and Semi-Static Context Injection
Modular Perception and Textual Abstraction
Integrated Visual-Language Reasoning
ABC
Human-Oriented Understanding
Emotional Grounding
Task Intent Formulation
Human Model Alignment
Generative and Agentic Interaction
Generative Social Communication
Collaborative Task Co-Creation
Proactive Agency
Persona Adaptation and Conversational Fluidity
Task-Oriented Planning and Execution
Social Initiation
Embodied Social Expressiveness
Creative Storytelling and Social Engagement
Anticipatory Assistance
Iterative Optimization and Alignment
Longitudinal Personalization and Memory
Sustained Personalization
Episodic Memory Integration
Multi-Level Repair
Behavioral Repair in Task Execution
Emotional Repair in Social Interaction
Repair in Ethical and Normative Alignment
Design components and strategies
Modality
Text
Voice
Visual
Motion
Hybrid
Other
Morphology
Humanoid
Humanlike appearance
Functional
e.g. arm/ mobile robot
Other
desktop/ animal-like
Autonomy
Teleoperation
Remote Control robot
Semi
With human intervention
Full
Operate with- -out human
Study Methods and Evaluation Strategies
Study Methods
Laboratory Experiment
Field Deploment
Interview
Questionnaire
Technical Evaluation
Evaluation Strategies
Objective
Task Accuracy and Performance
Task Efficiency and Timing
LLM
86/100
LLM-Specific Performance
Subjective
Robot usability
Safety
Anthropomorphism
User's Perceptual and Relational Experience
Perceived Intelligence
Cognitive Load and Workload
Application
Social and Conversation Systems
Teaching and Education
Healthcare and Wellbeing
AR/VR-enabled Interaction
Domestic and Everyday Use
Public Space Service
Industrial Manufac-turing
Other Application
Challenges
• Reliability of LLM-driven Understanding
• Multimodal Perception of Emotional Intelligence
• Multimodal information sensing and alignment
• Equitable Engagement in Multi-User Scenarios
• Morphology-Aligned Social Intelligence
• Balancing Autonomy and Human Oversight
• Balancing Trust and Overtrust
• Safeguarding Privacy and Mitigating Safety Risks
• Stabilizing Personalization and Social Alignment
• Sustaining Long-Term Engagement
• Proactive Repair for Diverse Contexts

Figure 1: Visual abstract of our systematic review of LLM-driven HRI, summarizing the overall structure of our work.

*Both authors contributed equally to this research.
†Corresponding authors

CHI '26, Barcelona, Spain
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790920

Abstract

Advances in large language models (LLMs) are profoundly reshaping the field of human–robot interaction (HRI). While prior work has highlighted the technical potential of LLMs, few studies have systematically examined their human-centered impact (e.g., human-oriented understanding, user modeling, and levels of autonomy), making it difficult to consolidate emerging challenges in LLM-driven HRI systems. Therefore, we conducted a systematic literature search following the PRISMA guideline, identifying 86 articles that met our inclusion criteria. Our findings reveal that: (1) LLMs are transforming the fundamentals of HRI by reshaping how robots sense context, generate socially grounded interactions, and maintain continuous alignment with human needs in embodied settings; and (2) current research is largely exploratory, with different studies focusing on different facets of LLM-driven HRI, resulting in wide-ranging choices of experimental setups, study methods, and evaluation metrics. Finally, we identify key design considerations and challenges, offering a coherent overview and guidelines for future research at the intersection of LLMs and HRI.

CCS Concepts

• Human-centered computing → Human computer interaction (HCI); **• Computer systems organization → Robotics**.

Keywords

human-robot interaction, large language models, human-centered robotics, systematic review, HRI, LLMs, LLM-HRI

ACM Reference Format:
Yufeng Wang, Yuan Xu, Anastasia Nikolova, Yuxuan Wang, Jianyu Wang, Chongyang Wang, and Xin Tong. 2026. How Do We Research Human-Robot Interaction in the Age of Large Language Models? A Systematic Review. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 28 pages. https://doi.org/10.1145/3772318.3790920

1 Introduction

Human–robot interaction (HRI) is a multidisciplinary field devoted to creating efficient, safe, and comfortable ways for people to collaborate with robots [1, 13]. Its enduring goal is to facilitate natural and effective interactions between humans and robots [45, 123]. However, the field has persistently faced significant challenges in achieving this, particularly in enabling robots to adapt to unexpected situations or unpredictable human behaviors in real-world, synchronous environments [31, 131].

The rapid advancement of large language models (LLMs) has introduced a transformative potential to address these challenges. To elaborate, LLMs enable robots to acquire, reason, and apply knowledge in physically grounded and socially environments [110] by enhancing their in-context learning [21], commonsense reasoning [151], and chain-of-thought capabilities [165]. Consequently, LLMs are reshaping the HRI landscape, powering innovations in emotional responses [99], adaptive task planning [192], context-aware tutoring [139, 149], and personalized care [44, 71, 86]. This integration constitutes a promising research frontier to advance embodied intelligence and enable more seamless collaboration [88].

Given the transformative impact of LLMs, we posit that it is a critical juncture to systematically review and synthesize this rapidly evolving landscape. The urgency of this synthesis is underscored by our preliminary analysis of publication trends in the ACM digital library (DL) (2015–2025), which reveals that the integration of LLMs into HRI has become a burgeoning and vital research domain since 2021 (see Figure 2). Furthermore, interdisciplinary human–computer interaction (HCI) methodologies facilitate the design and evaluation of technologies from a human-centered perspective, adhering to guidelines like ISO 9241-210:2019 [38, 61]. However, our survey of the current literature highlights a critical gap: while existing reviews predominantly focus on technical implementation, such as model robustness and architectural advancements [185], they often overlook the human-centered considerations that are foundational to the HRI field, including human-oriented understanding [41], user modeling [129], and levels of autonomy [45]. Moreover, discussions in current HRI studies remain fragmented, lacking a cohesive synthesis that aligns these technological leaps with established human-centered perspectives [117].

To address these gaps and provide a structured overview of this vital field, we conducted a systematic review following the PRISMA guidelines [108]. Our review specifically focuses on LLM-driven HRI studies from the past five years that involve practical interaction scenarios and examine the evolving role of LLMs in shaping the interaction lifecycle, resulting in 86 papers for in-depth analysis. Our investigation is structured around the following research questions: (1) **RQ1:** How do LLMs transform the foundational capabilities of HRI? (Section 4 proposes the Sense–Interaction–Alignment framework); (2) **RQ2:** How are LLMs integrated into HRI system design? (Section 5 discusses design components and strategies); (3) **RQ3:** How can LLM-driven HRI systems be evaluated? (Section 6 explores study methods and evaluation strategies); and (4) **RQ4:** What are the opportunities and challenges for future research? (Section 7 and 8 discuss applications and challenges respectively).

As LLMs inject new vitality into HRI, we systematically synthesize the current landscape of LLM-driven robotic systems to offer researchers a holistic overview connecting foundational capabilities, system design, and evaluation, while highlighting emergent challenges to inform future directions. Our analysis reveals that existing research has increasingly organized around a Sense-Interaction-Alignment paradigm, marking a key shift from rigid, task-specific pipelines toward adaptive, socially aware, and iteratively optimizable embodied intelligence. We further identify substantial heterogeneity in how LLMs are integrated across robot autonomy, interaction modalities, and physical embodiments, as well as a dual-focus evaluation trend that jointly considers objective task performance and subjective human experience. In summary, this paper presents the following contributions:

- A synthesis of LLM-driven HRI studies, demonstrating how LLMs enable contextual sensing, generative interaction, and adaptive alignment in embodied settings, and providing insights to support researchers in navigating the evolving landscape of LLM-HRI integration.
- A proposed systematic taxonomy, identifying the core areas emphasized in LLM-era HRI research and offering a structured categorization of studies across nine key dimensions.

- Identification of emerging challenges for future research directions, highlighting key design considerations, such as ensuring the reliability of LLM-driven understanding, maintaining appropriate levels of user trust, and achieving robust multimodal grounding in dynamic environments.
- An open-access, searchable online database containing the 86 included studies[1]. The platform allows users to browse the literature, perform retrieval, and interact with visualized charts, thereby supporting transparency and reproducibility.

2 Scope and Related Work

2.1 Scope

This section defines the key terms used in this paper to establish a common ground for our discussion and clarify the scope of this study.

2.1.1 *Robot.* A robot is conventionally defined as a physical agent that autonomously perceives its environment via sensors and acts through effectors [138], generating behaviors to accomplish one or more tasks [7]. According to ISO 8373:2021 [62], this is further specified as a "programmed actuated mechanism with a degree of autonomy to perform locomotion, manipulation, or positioning." Complementing this technical framing, HCI and HRI work often conceives robots as an umbrella term describing a miscellaneous collection of (semi-)automated devices with various capabilities and appearances [45], ranging from traditional industrial robots to simulated agents and actuated user interfaces [78, 104, 150]. Despite this inclusive conceptualization, several studies further emphasize that a robot's physical presence and embodied action are its most critical properties, as they enable the system to act in, sense, and reshape human environments [111, 166, 167]. In this paper, we adopt a precise definition of robots as entities possessing a physical embodiment or a simulated physical presence (e.g., in VR or AR environments) [41]. This includes systems such as robotic arms [59, 68, 94] and humanoid robots [49, 112] that interact with the physical environment (pHRI) through sensors (e.g., cameras, LiDAR) and actuators (e.g., motors, joints). While we acknowledge that purely disembodied agents, such as chatbots, play a pivotal role in natural language processing research [137, 169], the primary focus of this work is to explore the unique implications of embodiment. Therefore, to maintain this research focus and better investigate what we consider the essential characteristics of HRI [29, 30], we deliberately exclude non-embodied agents.

2.1.2 *Human-Robot Interaction.* As a multidisciplinary research domain, HRI integrates perspectives from HCI, psychology, and sociology [29, 66, 146], with a primary emphasis on Human-Centered Design (HCD) [19, 100, 174, 189]. To delineate the scope of HRI, we adopt the classification framework proposed by Sheridan [134], which identifies four major interaction types: (1) supervisory control; (2) teleoperation; (3) automated vehicles; and (4) social interaction, as illustrated in Figure 3. To ensure empirical grounding, HRI research relies on user studies, in which you measure how users respond to variations of the robot, the interaction itself, or the context of the interaction [13, 79]. Typical approaches include Wizard-of-Oz studies [28, 122], structured interviews [156], questionnaires [125], and field deployments [22].

2.1.3 *Large Language Model.* Large language models are natural language models that refer to Transformer-based architectures comprising billions of parameters [191], with notable examples including GPT-3 [40], Grok 3 [172], and LLaMA [154], as well as recent multi-modal LLMs (MLLMs) such as GPT-4 [106], GPT-5 [105], and Gemini [43]. Although these architectures were originally designed for Natural Language Processing (NLP) tasks [74], the field has rapidly expanded into the multimodal domain. This evolution has driven the emergence of vision-language models (VLMs) and MLLMs [179], which integrate textual, visual, and sometimes auditory modalities to enable richer forms of perception, reasoning, and interaction. Given this rapid expansion and the need to ensure our review remains sufficiently forward-looking, we adopt a broad definition of LLMs that encompasses pre-trained language models (PLMs) [52, 118, 194], as well as cutting-edge VLMs [132, 186], and MLLMs [84, 195].

2.2 Related Work

To provide a structured overview of related studies and their contributions, we summarize key prior works in Appendix B, categorizing them by type, focus, and main contributions.

With the rapid development of LLMs, an increasing number of surveys and reviews have emerged in this domain, addressing different aspects of robotics, such as robotic systems [39, 160, 181], robotic intelligence [64, 72], robot autonomy [90, 161], multi-agents [51, 85, 173], and embodied AI [88, 128]. For example, Zeng et al. [181] survey mainstream LLMs and analyze interaction as related technologies, with a focus on game-based and language-based HRI. Wang et al. [160] and Liu et al.[90] situate HRI in the robot task category, offering analysis in areas such as natural language interaction, task planning, and interaction experience. Xi et al. [173] broadly categorize the interactions of LLM-based agents into two types: cooperative interaction and adversarial interaction. Collectively, these works provide domain-specific analyses that enrich the broader understanding of LLM-driven robotics and offer conceptual perspectives that help situate and contextualize our focus on HRI. There are some studies initially focusing on LLMs in a HCI perspective. Among these works, Zhang et al.[185] provide the first review of how LLMs enhance HRI across inquiry answering, commonsense, and instruction following, while highlighting key challenges in safety, context understanding, and scalability. Shi et al.[135] demonstrate that, within socially assistive robotics, LLMs enable natural dialogue, multimodal user understanding, and policy synthesis. Zou et al. [196] propose a general taxonomy of LLM-driven human–agent systems, revealing that incorporation of environment and profiling, human feedback, interaction types, orchestration, and communication enhances system performance, reliability and safety. Atuhurra's meta-study of 250 HRI papers [8] corroborates these technological benefits while offering a critical counterpoint: the LLM substitution for traditional speech, intent, and perception modules that enriches knowledge, reasoning, and personalization in social robots simultaneously intensifies algorithmic bias, data-leakage risks, and computational overhead.

[1] https://llms-hri.github.io/

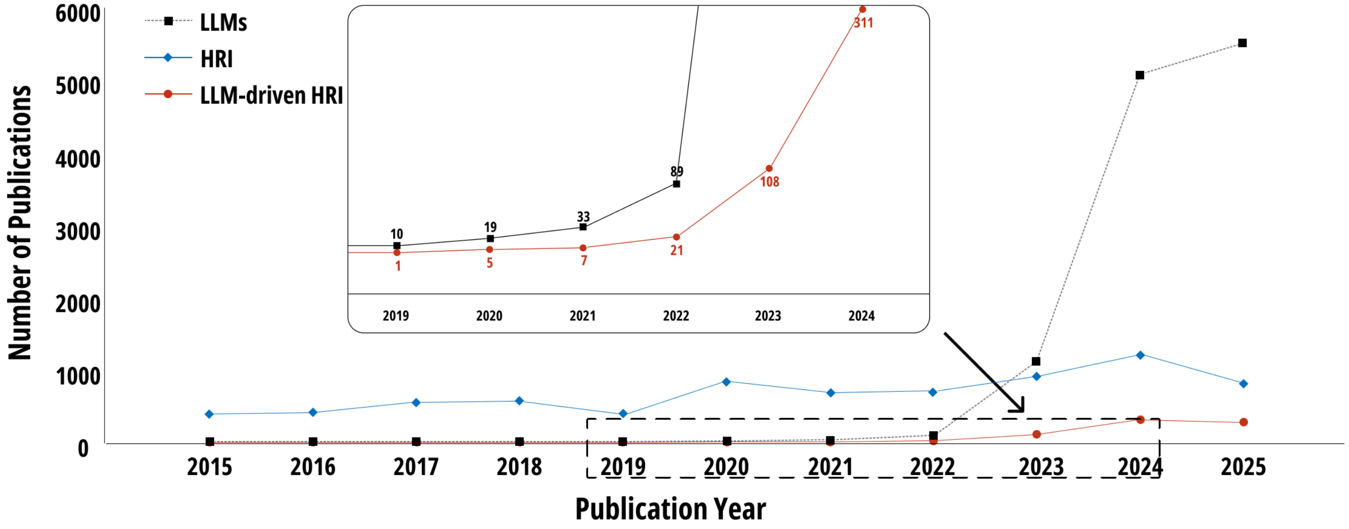


Figure 2: Publication trends of three research domains (LLMs, HRI, LLM-driven HRI) in the ACM digital library from 2015 to 2025 (details of search keywords are provided in Appendix A).

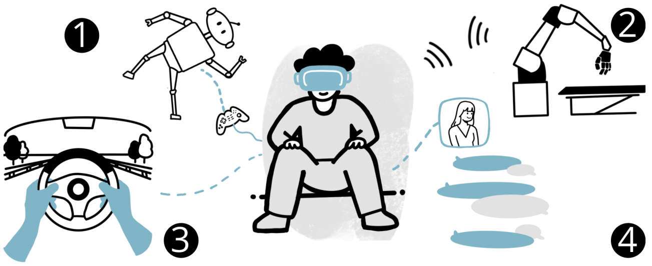


Figure 3: Illustration of the four core HRI types in Sheridan's classification framework: (1) supervisory control; (2) teleoperation; (3) automated vehicles; (4) social interaction.

While these prior studies have analyzed and evaluated the role of LLMs, and revealed possible directions in the post-LLM era, their scope remains limited. First, existing works often concentrate on the technical potential and performance of LLMs, such as advancements in core model architectures [185], improvements in training datasets [39, 90, 161], and fine-tuning paradigms [85]. Second, most works lack systematic and transparent procedures for literature identification, inclusion, and analysis. Further, LLMs have quickly expanded from text-only models to increasingly capable multimodal systems, and these developments are poised to introduce new opportunities, challenges, and design considerations for HRI that require timely integration and synthesis. To fill the gap, we provide the first systematic review of LLM-driven HRI. We adopt a rigorous and transparent methodology to synthesize how HRI is being reconceptualized in the era of LLMs and to outline potential directions for future LLM-driven HRI research.

3 Methodology

This systematic review was conducted and structured following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [108].

3.1 Search Strategy

A complete list of all databases, search results, and inclusion counts is provided in the Appendix C. In this section, we focus on detailing the formulation of our search queries and the rationale behind our database selection.

3.1.1 Search Query Formulation. After surveying a substantial body of literature, we established the scope of our review as outlined in **Section 2.1**. To construct a robust search strategy, we first conducted an exploratory search on Google Scholar using the query: *robot AND (llm OR vlm OR mllm OR gpt) AND hri*. However, this initial attempt revealed two limitations: (1) a large number of non–peer-reviewed arXiv manuscripts, and (2) considerable noise caused by acronym-based searches. For example, LLM is also commonly used to refer to the Master of Laws degree [152]. To mitigate these issues, we turned to ACM DL [2], whose technical focus helped reduce such ambiguity.

From this initial query, we identified 157 papers, which our research team divided for close reading. After a week of analysis and group discussion, several insights emerged that helped refine and supplement our scope. In this process, we found that some papers only mentioned LLMs briefly in discussion or future work sections, without integrating or evaluating them as core system components [10, 18, 70, 120]. Others identified limitations of existing LLMs to emphasize the strengths of their own proposed methods [20, 97, 171]. These observations helped establish a stricter inclusion criterion: we focus exclusively on studies that integrate,

[2] The search was conducted on July 22, 2025.

evaluate, or operationalize LLMs as technical components for addressing HRI problems.

During the close reading, we noticed that many HRI studies also used the term Human–Robot Collaboration (HRC) to place greater emphasis on collaborative scenarios [46, 55, 119, 190]. As we know, the robotics community commonly regards physical Human–Robot Collaboration (pHRC) as falling within the broader scope of pHRI [2, 3, 102]. Further, recent HRI reviews place particular emphasis on collaborative interaction [23] and even include HRC-related terms in their search strategies to ensure adequate coverage [163]. Therefore, to avoid omissions, we accordingly expanded our terminology. A similar refinement was needed for robot-related keywords. Our decision was guided by two considerations. First, because our review adopts a user-centered perspective, social robots and humanoid robots play a particularly important role. Social robots represent highly interaction-oriented platforms [116], and humanoid robots tend to afford more natural and embodied user-centered interaction [147]. Second, our preliminary screening revealed that a large proportion of retrieved studies already centered on social and humanoid robots. Although the general term "robot”could technically encompass these categories, we chose—after consulting domain experts—to include "social robot”and "humanoid robot”in our search terms to better reflect our human-centered orientation and ensure more comprehensive coverage of interaction-focused HRI work.

These iterative refinements led to our final search query, which incorporates full-form terms, representative model names, and key concept variants: *(“large language model”OR LLM OR ChatGPT OR GPT-3 OR GPT-4) AND (robot OR robotics OR “social robot”OR “humanoid robot") AND (“human-robot interaction”OR HRI OR “human robot collaboration”OR HRC).* We intentionally did not include general keywords such as "artificial intelligence", since our aim is to identify work that specifically involves LLMs rather than the broader AI literature (as seen in prior reviews of human–AI interaction [33, 98, 193]). Similarly, we did not enumerate additional model names such as Gemini or LLaMA, and instead opted for broad conceptual terms (e.g., large language model) to ensure comprehensive field coverage without overfitting to specific model families [152].

3.1.2 Database Selection. Given the interdisciplinary nature of our review, which integrates perspectives from both HCI and robotics, we examined high-impact publication venues in the two fields using Google Scholar’s “Top Publications” lists for HCI[3] and Robotics[4]. We identified the ACM DL as the primary repository for HCI research and IEEE Xplore as the leading database for robotics publications. Therefore, we selected ACM DL and IEEE Xplore as our main databases. To broaden the scope and include interdisciplinary perspectives crucial to HRI, the search was expanded following consultation with subject matter experts. This expansion included four high-impact, cross-disciplinary publication venues: Nature, Science Robotics, Computers in Human Behavior, and the International Journal of Social Robotics.

3.2 Screening and Selection

3.2.1 Inclusion and Exclusion Criteria. The screening and selection process was conducted in multiple stages to systematically refine the initial pool of literature down to a final, relevant corpus. The process was guided by a predefined set of inclusion and exclusion criteria, applied consistently by two independent reviewers.

To be included in the final review, a study had to meet several key criteria. Specifically, the publication was required to be a full-length, peer-reviewed research article in English that presented an empirical study on the integration of an LLM with a robotic system for an HRI application **(Section 2.1.2)**. A crucial inclusion criterion was the presence of an embodied agent, which we defined as either a physical robot or a high-fidelity simulated proxy that allows for spatial and interactive presence, such as avatars in VR or AR environments **(Section 2.1.1)**. Conversely, studies were excluded based on several factors. We excluded non-empirical works such as literature reviews, surveys, conceptual frameworks, and theoretical discussions that lacked a direct robotic application. Articles were also removed if LLMs were only mentioned superficially (e.g., in background or future work sections) rather than being a core component of the reported system **(Section 2.1.3)**. Furthermore, we excluded studies that focused purely on technical improvements or system architecture without involving a user interaction or evaluation process **(Section 2.1.1)**. Research centered on purely disembodied agents, like text-based chatbots, was not included **(Section 2.1.1)**, nor were non-full papers such as workshop articles, technical reports, or abstracts.

3.2.2 Resolution of Disagreements. During screening and coding, disagreements between the two primary reviewers were resolved through a staged process. The coders first discussed each discrepancy to reach consensus. If agreement could not be achieved, they discussed with a third reviewer to ensure reliability.

The first point of disagreement concerned papers from Scientific Reports. Although labeled as a "report”, this venue predominantly publishes full empirical studies rather than non-refereed reports. After discussion, we included two papers that satisfied the criteria [53, 92]. In total, sixteen papers produced inclusion–exclusion disagreements. For example, Hsu et al. [56] present a three-year robot study with people living with dementia, shifting from WoZ control to GPT-driven autonomy. It delivers actionable "research as care”guidelines, bridging technical HRI design with humanistic care for vulnerable populations. Therefore, we include it for unique insights on adapting LLMs for translating care ethics into HRI practices. Conversely, papers such as Promises [50] were excluded. Although participants interacted with a commercially available LLM-enabled robot (e.g., Moxie), the study did not analyze or evaluate the LLM components themselves, nor were LLMs methodologically central to the research design. Through adjudication, eleven papers were ultimately included, and five were excluded.

Given that the disagreement rate was higher than expected, we additionally conducted a false-negative check to assess whether any relevant papers might have been mistakenly excluded. We randomly sampled 100 papers from the 306 studies that had been explicitly excluded during screening [5]. Two coders independently

[3] https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_humancomputerinteraction

[4] https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_robotics

[5] Sampling was performed using a random integer generator with uniform sampling without replacement.

re-evaluated these papers against the inclusion criteria. Only three borderline papers were identified as potentially relevant; however, closer examination showed that one primarily centered on dataset construction [82], while the other two relied on video or image demonstration and therefore lacked any real interaction with a physical robot [80, 130].

During the first round of open coding, we noticed that certain categories produced a higher number of disagreements. For instance, categories such as Contextual Perception and Understanding and Evaluation Metrics were sometimes defined implicitly by authors, which made interpretation less consistent across coders. In the case of usability, for example, some papers reported standardized instruments such as the system usability scale [32, 42, 73, 164], whereas others embedded usability-related assessments within broader interview or questionnaire responses [27, 145]. To quantify the degree of coder agreement, we computed inter-rater consistency (IRR) using Cohen's kappa coefficient [26]. By contrast, highly observable categories such as modality (IRR = 0.768), robot morphology (IRR = 0.894), and application domain (IRR = 0.904) demonstrated strong agreement due to their concrete and surface-level characteristics. To address lower-agreement categories, we held a calibration meeting and refined the operational definitions before conducting a second coding pass. This resulted in substantially improved IRR scores. Finally, all the categories, subcodes, and IRR values, is provided in Table 1.

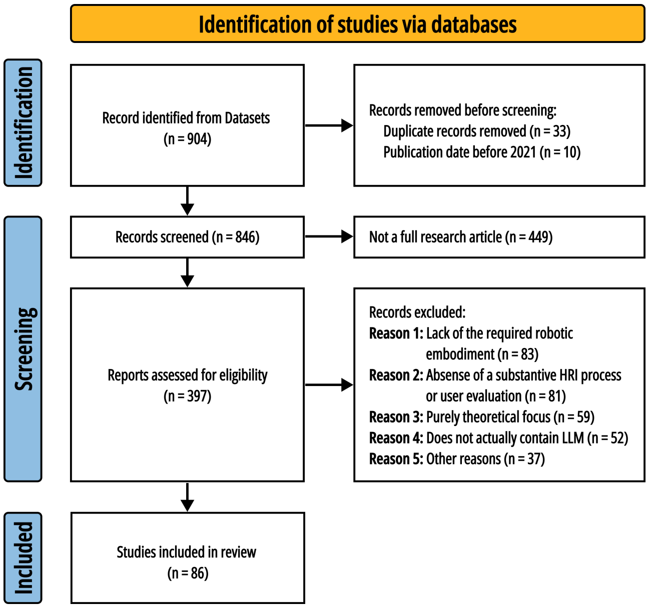

Identification of studies via databases
Identification
Record identified from Datasets (n = 904)
Records removed before screening:
Duplicate records removed (n = 33)
Publication date before 2021 (n = 10)
Screening
Records screened (n = 846)
Not a full research article (n = 449)
Reports assessed for eligibility (n = 397)
Records excluded:
Reason 1: Lack of the required robotic embodiment (n = 83)
Reason 2: Absense of a substantive HRI process or user evaluation (n = 81)
Reason 3: Purely theoretical focus (n = 59)
Reason 4: Does not actually contain LLM (n = 52)
Reason 5: Other reasons (n = 37)
Included
Studies included in review (n = 86)

Figure 4: PRISMA flow diagram outlining the literature screening and inclusion process for this systematic review.

3.3 Final Corpus

3.3.1 Paper Selection Process. A complete PRISMA flow diagram illustrating this entire process is provided in Figure 4. The initial database search returned a total of 904 records. After removing 33 duplicates and 10 articles published before 2021, 846 unique records remained for the initial screening stage. During the title and abstract screening, 449 records were excluded, primarily because they were not full research articles, leaving 397 articles for a full-text eligibility assessment. In the final full-text review stage, each of the 397 articles was read in its entirety. This rigorous assessment resulted in the exclusion of a further 311 articles. The primary reasons for exclusion were a lack of the required robotic embodiment (83 studies), the absence of a substantive HRI process or user evaluation (81 studies), and a purely theoretical focus, such as being a survey or conceptual framework (59 studies). Additionally, 52 studies did not substantively integrate an LLM in their system, and a further 37 were excluded for other reasons, such as having a primary goal of dataset collection rather than HRI investigation, or meeting multiple exclusion criteria simultaneously. This multi-stage selection process resulted in a final corpus of 86 studies, which form the foundation of our systematic analysis.

3.3.2 Overview of Included Papers. The distribution of publication venues and years for the papers included in this review is presented in Figure 5. An analysis of the publication venues reveals that the included papers are disseminated across premier conferences and journals in the intersecting fields, which underscores the highly interdisciplinary nature of research on LLM-driven HRI research. Notably, publications from the ACM/IEEE International Conference on Human-Robot Interaction and the ACM CHI Conference on Human Factors in Computing Systems collectively account for the largest proportion of the literature included in our review. Regarding the publication time in Figure 5.b, the number of included publications demonstrates a marked increase over time, with a significant concentration of works appearing in 2024 and 2025.

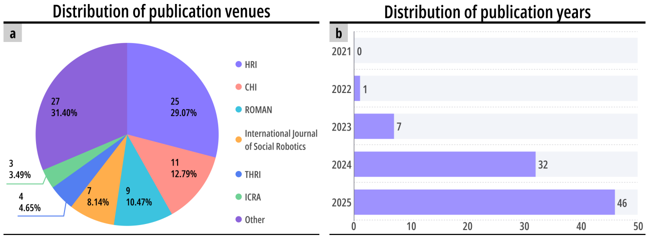


Figure 5: Overview of publication venues and years: (a) Distribution of included papers by venue. Venues contributing fewer than two included papers were not reported individually and are grouped under "Other" to ensure a meaningful representation. (b) Annual numbers of included papers.

4 Large Language Models in Human-Robot Interaction

In this section, we aim to answer **RQ1**. Overall, LLMs have endowed robots with a formidable cognitive foundation, characterized by zero-shot capabilities, complex reasoning, and in-context learning [21, 151, 165]. However, transposing these capabilities from disembodied text processing to the physical reality of HRI involves multiple transformations. To structure this transition, we adapt the classical "Sense-Plan-Act" [101] and "Reason + Act" [177] paradigms into a "Sense-Interaction-Alignment" framework, as

Table 1: The final codebook with 9 code categories and 60 subcodes; the average IRR is computed across the subcodes for each code. Multiple refers to "multiple codes can apply". In "Methodology", "Other Methods" occurred too infrequently to be included in the quantitative analysis.

Category	Codes	Mean IRR	Multiple
Contextual Perception and Understanding	Static and Semi-Static Context Injection; Modular Perception and Textual Abstraction; Integrated Visual-Language Reasoning; Emotional Grounding; Task Intent Formulation; Human Model Alignment	0.720 (SD=0.108)	Yes
Generative and Agentic Interaction	Persona Adaptation and Conversational Fluidity; Embodied Social Expressiveness; Task-Oriented Planning and Execution; Creative Storytelling and Social Engagement; Social Initiation; Anticipatory Assistance	0.796 (SD=0.073)	Yes
Iterative Optimization and Alignment	Sustained Personalization; Episodic Memory Integration; Behavioral Repair in Task Execution; Emotional Repair in Social Interaction; Repair in Ethical and Normative Alignment	0.684 (SD=0.070)	Yes
Modality and Interaction Channels	Text; Voice; Visuals; Motion; Hybrid; Tangible and Haptic Interaction; Proximity	0.768 (SD=0.070)	Yes
Morphology	Humanoid; Functional; Zoomorphic; Desktop Companions; AR/VR	0.894	No
Autonomy	Full Autonomy; Semi-Autonomy; Teleoperation	0.846	No
Methodology	Laboratory Experiment; Field Deployments; Interviews; Questionnaires; Technical Evaluation; Other Methods (WoZ; Case Study; Simulation; Co-Design Workshops; BodyStorming; Think-Aloud Protocols)	0.684 (SD=0.172)	Yes
Evaluation Metrics	Task Efficiency and Timing; Task Accuracy and Performance; LLM-Specific Performance; User's Perceptual and Relational Experience; Perceived Intelligence; Anthropomorphism; Usability; Safety; Cognitive Load and Workload	0.832 (SD=0.100)	Yes
Application	Social and Conversational Systems; Healthcare and Wellbeing; Domestic and Everyday Use; Teaching and Education; Industrial Manufacturing; AR/VR-enabled Interactions; Public Spaces Service; Other	0.904	No

seen in Figure 6. In the **Sense** phase, we argue that robotic systems ground abstract LLM capabilities within specific physical and social contexts to achieve true embodied intelligence, distinguishing them from disembodied AI agents [60, 89]. Second, we reframe the traditional notion of "Action" as **Interaction**, positing that LLM-driven behaviors are not solitary executions but proactive, multi-agent collaborations. Finally, we introduce **Alignment** as the critical adaptive phase, where the "Human in the Loop" (HITL) [38] necessitates continuous optimization through personalization and repair mechanisms to ensure robot behaviors remain congruent with human needs and social norms over time.

4.1 Contextual Perception and Understanding

In the **Sense** phase, LLMs transcend raw data processing to construct a semantic understanding of the environment. This section elucidates the transformation of sensory inputs into actionable cognitive contexts [81], progressing from the perception of physical surroundings to the understanding of complex social and affective dynamics.

4.1.1 Multimodal Physical Perception. The prerequisite for embodied intelligence is the ability to perceive and semantically interpret physical space. Unlike traditional robotics, which relies on rigid metric representations, LLMs facilitate the semantic parsing of environmental cues. We categorize these grounding strategies by their degree of semantic integration.

- ***Static and Semi-Static Context Injection.*** Primitive approaches rely on manual context injection, embedding environmental constraints directly into system prompts [59, 107, 145, 157, 176, 184] or utilizing pre-configured semantic maps [197]. These methods often incorporate granular object details, such as the dimensions of laboratory equipment [68] and spatial relationships [54, 65]. While effective in controlled settings, their reliance on "semi-static information" creates a bottleneck, limiting adaptation in dynamic or unmapped environments where real-time updates are essential [176].

- ***Modular Perception and Textual Abstraction.*** To overcome the limitations of static prompts, researchers have adopted dynamic Sensor-to-Text pipelines [47, 53, 76]. Systems like ARECA translate quantitative metrics (e.g., temperature, location) into narrative descriptions [24]. Concurrently, modular vision algorithms (e.g., YOLO, SAM) extract object labels [17, 76], while automatic speech recognition converts audio into text [69, 77, 113]. These text-based percepts serve as intermediate abstractions, enabling LLMs to perform high-level reasoning—such as determining object saliency from feature lists—without processing raw visual data directly [32, 37].

- ***Integrated Visual-Language Reasoning.*** Advanced implementations are increasingly replacing intermediate textual abstractions with VLMs for direct scene interpretation [15]. This shift enables real-time interaction captioning [42, 47, 94] and the integration of dynamic knowledge graphs for affordance planning [109]. Hybrid approaches further bridge precision and reasoning by combining fiducial markers (e.g., ArUco) with MLLMs to support complex behaviors like curiosity-driven exploration [81]. However, the reliance on converting multimodal inputs into a unified semantic space remains a significant challenge for achieving deep, lossless multimodal fusion. We will discuss this further in subsequent sections.

4.1.2 Human-Oriented Understanding. Beyond physical environment, effective HRI requires navigating the nuanced landscape of social interaction. Here, LLMs enable robots to shift from determining "what is there" to deciphering "who is there" and "why," processing invisible states such as emotion, intent, and social model.

- ***Emotional Grounding.*** Social understanding begins with the interpretation of emotional cues. Systems increasingly employ multimodal fusion, inferring affective states by combining facial detection with LLM-based textual analysis [17, 103, 140, 180] and incorporating temporal audio cues for greater accuracy [112, 143]. Recent research pushes towards "empathic grounding,"enabling robots to grasp complex nuances like nostalgia or implicit regret [5, 12, 48, 57]. This macro-level understanding, often spanning multiple dialogue turns, is critical for sensitive applications such as longitudinal well-being monitoring [6].

- ***Task Intent Formulation.*** Moving from emotion to pragmatics, LLMs facilitate the separation of explicit task requests from broader communicative intent [49, 114, 145]. For explicit commands, LLMs parse unstructured language into rigid specifications (e.g., task type, time, equipment) [15, 68, 164]. This capability extends to multimodal inputs, allowing robots to infer goals from sketches [42, 197] or body language [14, 83]. Furthermore, context-aware systems can predict task progression, such as detecting completed workflow steps [94, 153], or supporting creative intents for inclusive interactions with neurodiverse populations [4].

- ***Human Model Alignment.*** The highest level of social cognition involves internalizing implicit rules of engagement. At the conversational level, LLMs identify subtle shifts, such as intentions to change topics [15, 49]. More broadly, hybrid architectures enforce social compliance by integrating interpretation rules with prohibitions and obligations [32]. At a deeper level, LLMs provide a pathway to Theory-of-Mind capabilities, serving as Zero-Shot Human Models that simulate "what the human would think," thereby enabling robots to predict and align with complex social dynamics [157, 184].

4.2 Generative and Agentic Interaction

In the **Interaction** phase, LLMs fundamentally reshape HRI from rigid command-response loops to fluid, generative, and agentic collaborations. Unlike traditional systems constrained by pre-scripted behaviors, LLM-driven robots exhibit the capacity to generate novel social signals, co-create complex plans, and autonomously initiate interactions based on environmental context.

4.2.1 Generative Social Communication. LLMs empower robots to transcend static dialogue trees, enabling communication that is emotionally tailored, stylistically adaptive, and multimodally expressive [69, 145].

- ***Persona Adaptation and Conversational Fluidity.*** Effective social communication begins with linguistic adaptation, where robots dynamically modulate their personality rather than relying on generic responses [5, 103]. Researchers achieve this by conditioning models on psychological frameworks like the Big Five traits [12, 92, 182]. This allows robots to manifest distinct personas adapted to specific social roles—from "cheerfully ironic" tones that increase warmth [141] to self-disclosing styles suitable for health mediation [168] or emotional elicitation in elder care [57, 87]. Sustaining these personas requires moving beyond slot-filling systems to LLMs that process unstructured inputs for context-aware coherence [69, 145]. Modern systems thus focus on maintaining conversational flow by analyzing dialogue history and syntactic cues to predict precise turn-taking moments, ensuring social presence remains believable [124, 142].

- ***Embodied Social Expressiveness.*** Linguistic fluency alone is insufficient for embodied presence; the core challenge lies in synchronizing verbal output with non-verbal behaviors [188]. Moving beyond isolated motion generation, recent systems leverage LLMs to orchestrate holistic behavioral responses, simultaneously outputting verbal utterances, emotional states, and physical cues—such as head nodding—to facilitate empathic grounding [6, 93]. This coordination is refined by control strategies that align physical signals with linguistic intent, such as using gaze aversion to mark turn-taking boundaries [115, 142] or synchronizing facial expressions with utterance sentiment [5]. To support this expressivity without compromising real-time performance, architectures often hybridize generative LLM planning with rule-based execution, minimizing latency while maximizing social impact [34, 75, 81, 180].

4.2.2 Collaborative Task Co-Creation. LLMs transform robots from passive tools into active partners, with this collaborative paradigm evident in both goal-oriented tasks and open-ended, creative undertakings.

- ***Task-Oriented Planning and Execution.*** Collaboration in physical tasks has shifted towards shared autonomy, where agency is distributed between human and machine [27]. Systems like GenComUI visualize LLM-generated plans, allowing users to verify task flows before action [42]. During execution, frameworks like LILAC enable users to provide natural language corrections that update the robot's control space in real-time, facilitating the learning of complex manipulations from minimal demonstrations [27]. This co-authorship is further supported by multimodal tools bridging abstract intent and precise control via augmented reality and timeline adjustments [59, 68, 94].

- ***Creative Storytelling and Social Engagement.*** In open-ended domains, robots utilize LLMs as creativity scaffolds to foster engagement through joint imagination. Rather than passively delivering content, systems like Jibo actively co-construct stories with children by offering divergent narrative ideas [4, 35, 95]. This paradigm extends to therapeutic contexts, where robots generate personalized narratives dynamically adapted to the cognitive needs of older adults [17, 57], transforming social interaction into a co-creative and cognitively stimulating experience [5, 69, 133].

4.2.3 Proactive Agency. The emergence of social agency from LLMs allows robots to initiate actions based on inferred context, shifting the interaction dynamic from reactive execution to proactive engagement.

- ***Social Initiation.*** Research suggests that users prefer robots that proactively communicate their capabilities, enabling smoother cooperation [121]. Proactive robots bridge the gap between presence and interaction by autonomously identifying and approaching users in socially appropriate ways [12, 145, 155, 168, 176]. Beyond physical initiation, systems like SONAR exhibit "proactive social agency" by engaging in small talk during warm-up phases [32, 114], or employing "situation controlling" strategies to manage user expectations before interaction begins [121, 148]. However, balancing agency with user control is critical; while low-granularity adjustments align with user preferences, excessive granular control can undermine the perception of the robot as an intelligent social agent [183].

- ***Anticipatory Assistance.*** The agency further manifests through the anticipation of user needs. Robots leverage multimodal perception to proactively offer assistance or initiate context-aware conversations [67, 87, 148]. This extends to emotionally intelligent behaviors, such as taking the initiative to elicit positive states based on user personality [103], or employing curiosity-driven strategies to bridge knowledge gaps [81]. By autonomously investigating uncertainties, robots evolve into active partners that not only answer questions but strategically ask them to deepen mutual understanding [63]. To maintain alignment during such proactive behaviors, agents explicitly inform users of completed actions via dialogue cues [145] and actively seek clarification when facing ambiguity [197]. In the event of errors, structured feedback detailing the cause allows users to understand both the outcome and the underlying reasoning, fostering trust [81].

4.3 Iterative Optimization and Alignment

While sense and interaction provide the operational capability for behavior, they do not guarantee its long-term suitability. Generative interaction introduces risks of hallucination, misalignment, or drift. The **Alignment** phase addresses the critical feedback loops required to optimize interaction, ensuring it remains safe, ethical, and personalized over time.

4.3.1 Longitudinal Personalization and Memory. To achieve embodied intelligence, robots transition from episodic interactions to continuous relationships. This process relies on active personalization strategies supported by persistent memory repositories.

- ***Sustained Personalization.*** Personalization encompasses the long-term optimization of system parameters beyond immediate reactions [12, 46, 133]. While short-term adaptation handles real-time adjustments for safety [15] or dialogue pacing [9], true embodiment requires refining decision-making models over extensive periods. Addressing the data-inefficiency of traditional Reinforcement Learning, recent advancements leverage LLMs to accelerate policy tuning. For instance, ChatAdp uses ChatGPT to generate synthetic feedback, creating personalized training data with minimal human effort [148]. Additionally, frameworks like LAMS allow users to explicitly teach robots new logic through natural language [153], while systems like VITA evolve dedicated user models to align with personality traits over time [143, 162].

- ***Episodic Memory Integration.*** Long-term personalization is also founded on the retention of shared history [14, 53, 56, 103, 124]. Systems utilize sophisticated memory architectures, such as DSR graphs, to accumulate context across sessions, capturing preferred topics and cognitive patterns [17, 32]. This historical awareness enables dynamic profiling, allowing robots to tailor coaching strategies or speech rates based on past interactions [67, 96]. Operationalizing this involves live memory repositories where dual LLMs form a continuous feedback loop between perception, retrieval, and action [92]. Crucially, this process is personality-dependent; the robot's distinct character influences how memories are encoded and retrieved, guiding consistent and appropriate actions [103].

4.3.2 Multi-Level Repair. As personalization deepens, so does the need for defensive mechanisms to prevent misalignment. We categorize these mechanisms into three levels: task execution, social interaction, and ethical compliance.

- ***Behavioral Repair in Task Execution.*** When physical tasks fail, repair mechanisms address both control and logic. For real-time control, frameworks like LILAC allow users to refine the control space via natural language corrections [27]. To preempt errors, systems like UJI-Butler incorporate Human-in-the-Loop verification for planned actions [14, 182]. At the logic level, systems like RobotGPT utilize simulation-based corrector bots to iteratively analyze and fix runtime errors in LLM-generated code until successful execution is achieved [65].

- ***Emotional Repair in Social Interaction.*** In social contexts, breakdowns require strategies to mitigate frustration and preserve trust [11, 56]. Beyond error correction, robots navigate the dynamics of politeness [73, 96]. Research emphasizes the use of multimodal signals—such as visual cues or gestures—to resolve misunderstandings effectively [75, 93]. Longitudinal studies further suggest that repair strategies evolve with the user relationship, transitioning from generic apologies to complex empathic structures that acknowledge user feelings [11].

- ***Repair in Ethical and Normative Alignment.*** Finally, behaviors strictly adhere to societal norms and ethical guidelines. Architectures like SONAR enforce social appropriateness through formal rules regarding prohibitions and obligations. These systems may also employ online learning to adjust fuzzy definitions of norms, such as social distance, based on validated interaction data [32, 83]. Maintaining trust with vulnerable populations particularly requires transparency about limitations to prevent unrealistic expectations [87, 143, 188] and implementing guardrails for sensitive topics [96]. Balancing this deep personalization with privacy concerns remains a pivotal challenge in ethical alignment [112].

5 Design Components and Strategies

Building upon the preceding discussion of LLMs fundamental capabilities in HRI, this section addresses **RQ2**. Beyond exploring foundational technicalities, we specifically examine the transformative impact of LLMs across three pivotal dimensions: modality, morphology, and autonomy.

5.1 Modality and Interaction Channels

To design and conduct research in LLM-driven HRI, a critical early step is selecting the appropriate modalities, or channels, through which the human and robot will interact. The literature demonstrates a wide array of choices in Figure 7, ranging from fundamental text-based commands to complex, multi-sensory experiences.

5.1.1 Text. Text serves as the most direct and foundational modality for interacting with LLMs, utilized for both user input and robot output. This includes its use for **direct command and instruction**, where many systems rely on users providing textual commands through a console or editor [15, 36, 68, 77, 140, 148, 157, 176, 184, 197]. Furthermore, text is the core of **conversational interaction**

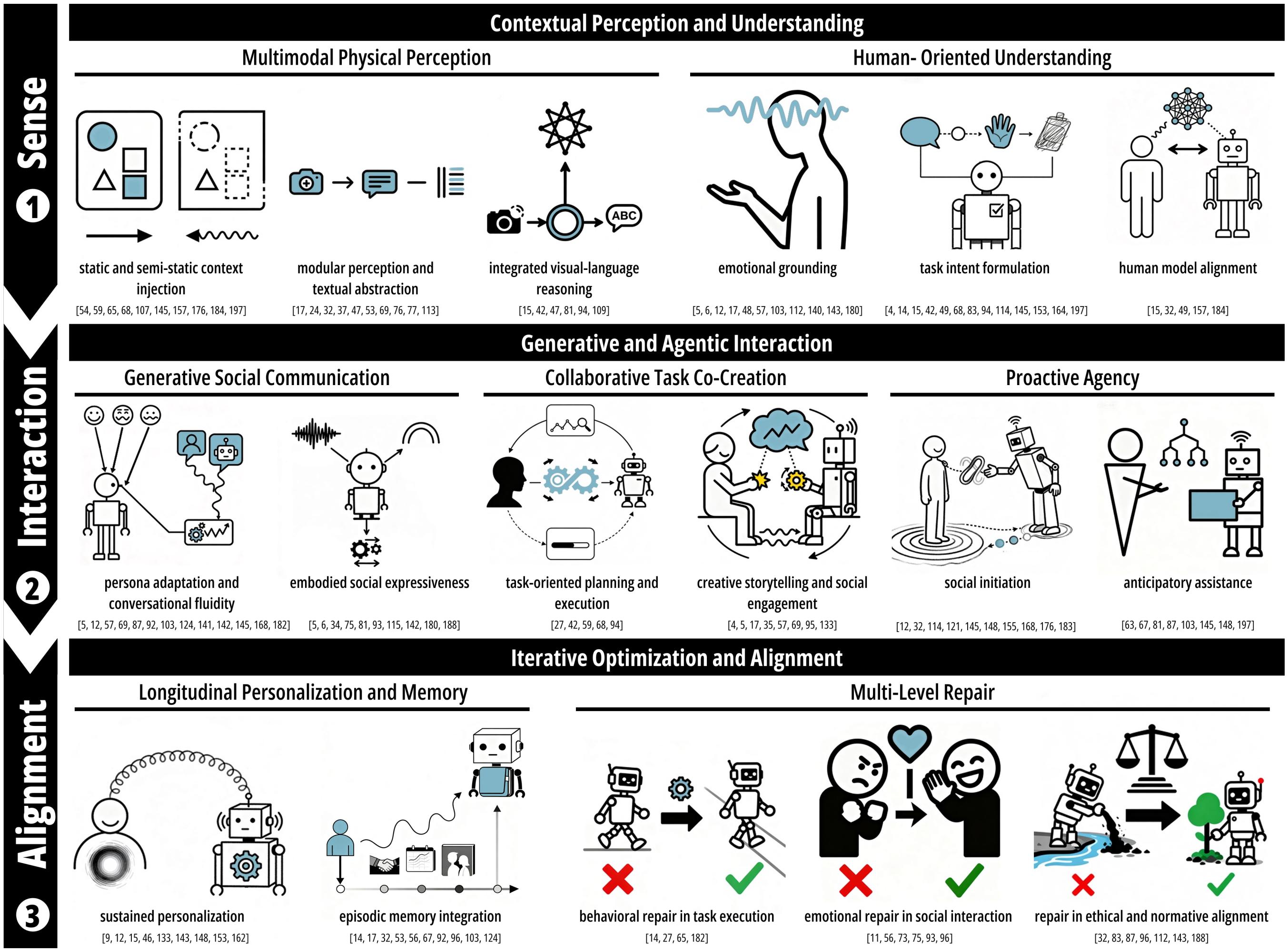


Figure 6: The proposed Sense-Interaction-Alignment framework for LLM-driven HRI research. This model adapts classical robotic paradigms to address the unique demands of embodied, social collaboration. It transitions from context grounding (Sense) and generative, multi-agent co-creation (Interaction), to continuous iterative optimization (Alignment).

beyond simple commands, as LLMs are used to generate conversational turns[158], rephrase dialogue for specific personas[115, 183], and formulate questions[63]. Finally, robots also employ text for **system output**, for instance, by displaying instructions or feedback on tablets and e-paper screens[25, 46].

5.1.2 Voice. Voice enables a wide range of spoken interactions from dialogues and small talk [114, 145] to live, unscripted conversations [73, 112], with the naturalness often enhanced by the underlying LLM [162]. The predominant **Speech I/O Pipeline** converts user speech to text via Automatic Speech Recognition (ASR) for the LLM, then vocalizes responses using Text-to-Speech (TTS) [15, 36, 53, 67, 69, 77, 81, 121, 127, 140]. Implementations leverage commercial services from OpenAI [4, 68, 175, 183], Google [5, 14, 54, 142], Microsoft [47, 83, 103], and Amazon [6, 9], alongside open-source [76] and platform-native engines [17, 124, 126, 158, 168]. A nascent trend **explores paralinguistic features** like pitch and speed [9, 12, 32] and synchronizes speech with physical embodiment [63, 133]. This allows robots to voice internal states [109], selectively address users [49], or manage turn-taking and recognition errors [142].

5.1.3 Visuals. The visual channel encompasses graphical interfaces, social cues, and environmental perception, with LLMs used to interpret visual input and generate visual output. Within this modality, robots **present information on screens**, from tablets [141] to dynamic maps with generative visual aids [42]. AR interfaces allow users to preview actions and define objects in the workspace [59, 68]. To **appear more social**, LLMs generate context-aware facial expressions [5, 133] and select gestures [54]. Abstract cues are also used, such as LED indicators that emulate breathing or manage turn-taking [24, 142]. Critically, **VLMs ground language in reality**. They allow robots to interpret scenes through dense captioning [47, 109] or understand human pose to provide context-aware feedback [159].

5.1.4 Motion. In physical interactions, motion is a primary modality where LLMs typically manage high-level task planning by translating abstract commands into action sequences, while low-level control relies on traditional robotics techniques. The application of motion can be broadly categorized into functional and expressive purposes. **Functional motion** involves goal-oriented physical actions to complete tasks. Examples include complex manipulation like pick-and-place, pouring, or grasping [65, 75, 76, 81], navigation to guide users or follow paths [121, 145, 197], and performing assembly procedures [114]. In parallel, **expressive motion** serves a communicative or social function to enhance interactional fidelity. It encompasses various forms of gestures [63, 69, 141], platform-specific expressive movements [4], and physical changes to convey system status, such as an air purifier's operational intensity [25].

5.1.5 Hybrid. Most sophisticated HRI systems are inherently multimodal, combining language with other channels to create more robust and intuitive user experiences [34, 81]. The effective orchestration of these channels is a key characteristic of advanced systems. A primary approach **synchronizes text-to-speech with non-verbal behaviors**. This includes generating corresponding facial expressions [5, 35, 112], body motion, and gaze [114, 162] to enhance social presence and convey intent [36, 56, 127, 140, 158]. Beyond expression, systems **fuse language with other modalities** for enhanced capability. This includes processing parallel voice commands with deictic gestures to resolve ambiguity [76], combining speech with sketches and sensor data for spatial tasks [68, 197], or linking dialogue to physical actions like navigation [15, 77, 145]. Some also integrate traditional interfaces like tablets [67].

5.1.6 Tangible and Haptic Interaction. This modality ranges from **direct physical guidance**, where a user manually moves a robot's arm to demonstrate a task [14], to **touch-based inputs on screens or sensitive surfaces** [35, 162]. Advanced applications of touch aim to simulate the experience of interacting with living entities, moving beyond simple functional feedback toward a more social, life-like feel [24, 25].

5.1.7 Proximity. The **spatial relationship** between a user and a robot is a subtle yet powerful social cue. Robots can manage conversational dynamics by adjusting their distance to users [49], and proximity can be explicitly modeled to interpret social situations [114]. The fundamental importance of this channel is underscored by the many systems designed for co-located, face-to-face interaction within a shared workspace [36, 76, 141, 184].

5.2 Morphology

Once the interaction channels are established, researchers further consider the physical morphology that serves as the carrier for these modalities. The physical embodiment of a robot represents a critical design consideration that can fundamentally shape the nature of interaction [41]. Figure 9 provides a survey of the diverse platforms utilized in the reviewed literature to support various interaction goals.

5.2.1 Humanoid. Humanoid robots (e.g., Pepper [12, 47, 49, 53, 63, 69, 124, 140, 141, 159], Nao [32, 48, 127, 182], Furhat [6, 9, 35, 67, 112, 127, 142]) remain the most prevalent due to their suitability for socially aligned interactions, with additional platforms such as QTrobot [57, 143], ARI [121], Navel [103], Geminoid F [126], and Mobi [92] are also frequently employed.

5.2.2 Functional. These robots are designed primarily for task performance rather than social embodiment. This group including mobile platforms (e.g., TurtleBot [14, 36], Segway [145]) and robotic arms [15, 27, 59, 65, 68, 73, 75, 83, 94, 107, 114, 153] dominate task-oriented applications such as manipulation.

5.2.3 Other. Some studies explore alternative morphologies. These include zoomorphic (animal-like) robots [17, 164], desktop companions like Haru [54, 133, 162, 180], and even vehicle-based agents for human-vehicle interaction research [144]. This diversity highlights the expanding design space for LLM-driven robots beyond traditional forms.

5.3 Levels of Autonomy

Moving from external form to internal logic, the design process further involves determining the appropriate level of autonomy. This dimension describes the extent to which a robot acts on its own accord and defines the distribution of control between the user and the LLM. It is important the optimal degree of independence often depends on the specific research context, the target application, and the nature of the task. Figure 8 illustrates the spectrum of autonomy, from direct control to more independent decision-making and execution.

5.3.1 Teleoperation. Teleoperation places a human in direct control of a robot, a method used to study user experience when autonomy is not yet feasible. In an assistive teleoperation model, LLMs can translate an operator's high-level commands into low-level robot actions to reduce cognitive load [153]. Even in these systems, human verification is common; for instance, an experimenter might approve LLM-generated responses before the robot speaks to ensure safety and appropriateness [58]. Additionally, teleoperation serves as a relatively important source for collecting data for robot imitation learning.

5.3.2 Semi-Autonomy. Semi-autonomous control serves as a practical strategy to mitigate system failures, including conversational breakdowns [49, 126], inappropriate LLM responses [53], and physical safety risks [73], acknowledging the continued need for human oversight when implicit cues are missed [57]. This approach is applied in diverse contexts: collaborative tasks [59, 68], creative content co-creation [4, 17, 83], and managing social dynamics like turn-taking and gaze [37, 54, 126]. Implementations typically involve a human-in-the-loop, sometimes via a Wizard-of-Oz setup [6, 126]. Other methods include direct human intervention to advance tasks [183], provide online corrections [27, 144], or curate AI-generated content [4] within a human-in-the-loop learning framework [14, 17].

5.3.3 Full Autonomy. Full autonomy aims to create robots that operate without direct human control [69], using LLMs to build end-to-end systems for complex social interactions [114, 145]. The objective is to produce socially aware partners [47] that generate their own real-time dialogue and behaviors, allowing researchers to study emergent user interactions [112, 121]. To enhance believability, these systems generate dynamic facial expressions [5] and context-sensitive gestures [12]. For complex tasks, they integrate multiple data sources using frameworks that combine LLMs with VLMs [81] and knowledge graphs [109]. Some systems even use reinforcement learning for dynamic adaptation [164]. Implementations are often hybrid architectures augmenting a central LLM with specialized components like computer vision for body tracking [158], models for turn-taking prediction [142], and standard speech I/O [67].

6 Study Methods and Evaluation Strategies

Regarding **RQ3**, this section systematically reviews the study methodologies in Figure 10 and evaluation strategies for HRI in the age of LLMs in Figure 11. Distinct from purely computational benchmarks, HRI research prioritizes human-centered evaluation, predominantly through user studies.

6.1 Methodology

6.1.1 Laboratory Experiment. Laboratory experiments remain a cornerstone for empirically evaluating LLM-driven HRI in controlled settings. These studies are essential for isolating the impact of LLM-driven capabilities on interaction quality and user perception. Researchers typically evaluate both **LLM-generated behaviors and dialogue** and **integrated end-to-end systems**. This involves using structured, task-based scenarios—such as collaborative manipulation [75], language learning [67], robot programming [68], or assistive tasks [76]—to gather performance metrics and subjective feedback. Dedicated labs allow for precise control over variables, facilitating high-quality data collection on how LLM-driven personalities [114], curiosity [81], or multimodal reasoning [47, 121, 197] affect user experience and task success.

6.1.2 Field Deployments. Field deployments test the robustness and adaptability of LLM-driven robots in real-world environments, moving beyond the constraints of the lab. These studies are crucial for understanding how systems perform over extended periods and with diverse user populations. Key applications include leveraging LLMs for **dynamic dialogue and contextual understanding** in unpredictable settings like classrooms [63] or public festivals [48, 53]. Another major focus is on **personalization and long-term adaptation**, particularly in home environments where robots support child wellbeing, act as companions [25, 133, 175], or assist older adults [56, 115]. Finally, deployments often evaluate **multimodal fusion and scenario-specific reasoning** in complex settings like care centers [17, 76] and university campuses [58, 164].

6.1.3 Interviews. Interviews are a key qualitative method for capturing nuanced human perceptions and expectations regarding LLM-driven robots. They are often used for **post-interaction evaluation**, like semi-structured formats allow researchers to gather detailed feedback on user experiences, preferences, and the reasoning behind them [5, 11, 42, 68, 81, 94, 153, 175, 188]. Interviews are also vital in **formative and scenario-based elicitation** during the early design stages, helping to align system concepts with user needs before implementation [56, 69, 109]. Furthermore, they are used to gather **expert and specialized feedback** from researchers or domain specialists on the broader implications of LLM-driven robotic systems [63, 133, 158].

6.1.4 Questionnaires. Questionnaires are a versatile tool for collecting quantitative, self-reported data at various stages of HRI research. **Pre-study questionnaires** gather demographic data and assess participants' prior experience or expectations with AI and robots [12, 32, 34, 133, 158, 162]. Their most common application is in **post-study questionnaires**, which measure subjective metrics like user enjoyment, perceived intelligence, and usability, capturing dimensions specific to LLM capabilities [11, 35, 47, 59, 68, 69, 112, 121, 143, 183]. Questionnaires also function as a **standalone research tool** for scalable data collection, such as gathering task instructions to fine-tune models [92, 164] or crowdsourcing evaluations of LLM-generated dialogue scripts [96].

6.1.5 Technical Evaluation. Technical evaluations of LLM-driven HRI systems assess functional correctness, fluency, and efficiency. This goes beyond traditional robotics to include **LLM-specific performance metrics** such as response latency, code execution success rates, and the semantic quality of generated language [5, 48, 63, 67, 94, 109, 121]. Since LLMs often act as a central hub, evaluations must also measure **multimodal interaction performance**, analyzing the fusion of language with perceptual data and comparing the integrated system against specialized baselines [32, 76, 114, 121]. Finally, **system-level efficiency and ablation testing** are used to measure computational overhead and isolate the LLM's specific contribution to performance gains, often validated with statistical tests [9, 47, 153, 164].

6.1.6 Other Methods. To address the unique challenges of evaluating LLM-based robots, researchers also employ a range of complementary methods. The **Wizard-of-Oz** technique remains prevalent for simulating advanced dialogue capabilities and mitigating model unreliability during user studies [6, 56, 58, 126, 144, 188]. **Case studies** and longitudinal deployments offer deep, contextual insights into long-term use [63, 109], while **simulations** enable safe and scalable testing in complex scenarios [164]. Lastly, **participatory methods** such as co-design workshops [54, 57], bodystorming [11], and think-aloud protocols [83] are critical for gathering formative feedback and aligning system design with user needs.

6.2 Evaluation Metrics

6.2.1 Objective. In the age of LLMs, objective evaluation in HRI combines traditional performance metrics with new measures assessing the models themselves.

- ***Task Efficiency and Timing.*** Metrics in this category quantify the speed and smoothness of the interaction. Researchers commonly measure **task completion time (TCT)** to assess overall efficiency [42, 76, 94, 114, 164]. Another key metric is **response latency**, which includes both the robot's general response time [32, 164] and the specific latency of LLM API calls, a factor critical for real-time turn-taking [115]. Additionally, studies evaluate the number of **dialogue turns** or interaction durations to gauge conversational compactness [42, 81, 112]. For instance, LLM integration has demonstrated the potential to reduce interaction time by up to 50% in certain multimodal tasks [76], highlighting its impact on efficiency.

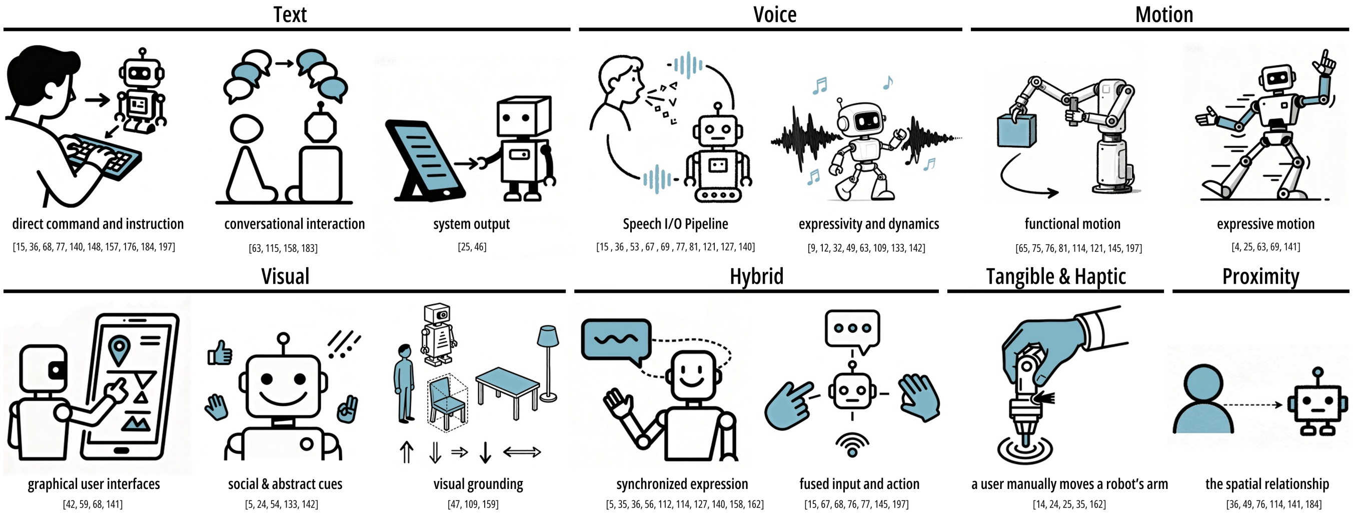


Figure 7: Interaction modalities, through which the human and robot will interact.

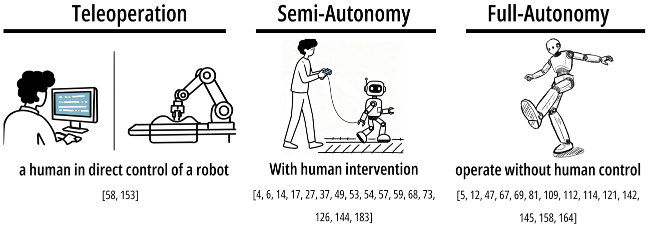


Figure 8: The increasing autonomy level from direct control to deciding everything and acting.

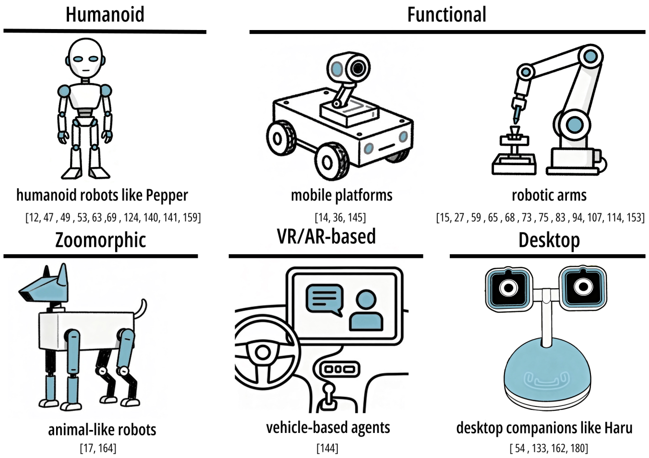


Figure 9: Different types of physical embodiment of a robot.

- ***Task Accuracy and Performance.*** This category evaluates the correctness and success of the HRI system. A primary metric is the **task completion or success rate**, often supplemented with error logs to identify failure modes [36, 59, 65, 68, 109, 145, 184]. Beyond binary success/failure, researchers employ more granular **accuracy scores** tailored to specific tasks. Examples include accuracy in question generation [63], semantic precision in understanding commands [37], accuracy in estimating user states or preferences [54, 126], and performance on benchmark datasets [92], with some systems achieving over 90% accuracy [176]. Other works also focus on performance in specialized domains like physics tutoring [77] or adaptive interfaces [15, 140, 148].

- ***LLM-Specific Performance.*** With LLMs as a core component, evaluation extends to the model's intrinsic performance. This involves measuring the LLM's **predictive accuracy** using standard machine learning metrics like recall, precision, and F1-scores for tasks such as turn-taking prediction [115] or theory of mind assessments [157]. The **quality of the generated output** is another crucial aspect, evaluated through metrics like Pylint scores for code generation [65], similarity to expert-designed behaviors [93], or accuracy in matching a target voice profile. These metrics provide direct insights into the LLM's reliability and capability, directly measuring its contribution to the overall system performance [180].

6.2.2 *Subjective.*

The integration of LLMs has profoundly reshaped the evaluation of subjective user experience in HRI. While established metrics for constructs like usability, safety, and perceived intelligence remain vital, they are now augmented by new dimensions that capture the complexities of LLM-driven interaction, such as conversational depth and relational quality.

- ***User's Perceptual and Relational Experience.*** The integration of LLMs has pivoted subjective HRI evaluation from static, task-oriented metrics toward a holistic assessment of the user's perceptual and relational experience. While foundational metrics like the Godspeed Questionnaire's *Likeability* subscale and general user **satisfaction** ratings remain prevalent [36, 69, 127, 133, 184], their scope has expanded. Satisfaction is now also judged by social dialogue quality [114, 143], emotional safety [57], and enjoyment from long-term interactions [121]. Similarly, **acceptance** now extends beyond reuse intention [121] to include attitudes toward

AI-generated content [54] and concerns about transparency and deception [162]. LLM-driven fluency also deepens **engagement**, transforming it into a measure of partnership and interaction quality [59]. This is often evaluated through the robot's ability to generate empathetic reactions that encourage user expression [5, 6, 34]. Finally, researchers increasingly assess **perceived robot qualities** like intelligence [47], empathy [180], curiosity [81], and creativity [35], focusing on the alignment between generated content and its physical, perceivable, and low-latency enactment [5, 12, 63, 93]. Longitudinal and multimodal methods are becoming essential to capture these dynamic aspects [54, 87, 133, 180].

- ***Perceived Intelligence.*** Perceived intelligence is a critical subjective metric in the LLM era, reflecting users' judgment of a robot's cognitive abilities. While traditional measures like the Godspeed questionnaire are still widely used to assess dimensions like competence and knowledge [42, 47, 183], LLMs introduce novel evaluative dimensions. Key among these are concerns about **factuality and accuracy**, as LLM hallucinations can negatively impact perceptions of intelligence [54]. Conversely, LLMs enable advanced social-cognitive abilities, such as **theory of mind** [92] and **emotional intelligence** [103], which expand the construct of perceived intelligence beyond task competence. Studies consistently show that the enhanced dialogue quality and responsiveness of LLM-powered robots lead to higher ratings in perceived intelligence [58, 126], an effect often underscored by direct user feedback such as "You're smart"[63].

- ***Anthropomorphism.*** The integration of LLMs has significantly reshaped the evaluation of anthropomorphism in HRI. Foundational dimensions like animacy, intelligence, and likeability, often assessed with the Godspeed questionnaire, remain relevant [42, 69, 127, 133, 140]. However, LLMs introduce more nuanced facets of human-likeness. Evaluation has expanded to include the robot's **conversational competence**, such as its ability to perform role-taking, maintain a consistent personality, and articulate values [53, 92, 126]. The perceived human-likeness is now deeply tied to the quality of dialogue, with compelling storytelling and personal reflections blurring the machine-human boundary [5, 24]. Furthermore, LLMs facilitate more complex and adaptive **personality simulations**, enhancing relatability [12, 92]. Nevertheless, user expectations are varied, underscoring the context-dependent nature of preferred anthropomorphism in LLM-driven HRI [32, 53].

- ***Usability.*** In the age of LLMs, HRI research frequently employs standard metrics like the System Usability Scale (SUS) to assess perceived usability across diverse applications [59, 68, 73, 81, 87, 94, 164, 175]. However, the unique capabilities of LLMs necessitate an expanded view of usability. Evaluations now also incorporate the quality of **dialogue-based interaction**, using instruments like the Chatbot Usability Scale [42]. Other critical LLM-specific dimensions include adherence to user preferences for **personalization** [183] and the provision of adjustable features that enhance **transparency and control** [109]. These aspects are crucial for building trust [14] and managing potential usability issues, such as users being distracted by overly engaging content [54].

- ***Safety.*** The integration of LLMs has broadened the scope of safety evaluation in HRI. While traditional measures of **perceived safety**—assessing users' comfort and security, often with the Godspeed Questionnaire—remain standard practice [15, 36, 42, 47, 67, 69, 127, 184], LLMs introduce critical new safety dimensions. A primary concern is **content safety**, as the generative nature of LLMs can produce outputs that are inappropriate or misaligned with user values [54]. Furthermore, the inherent **unpredictability and potential for logical failures** in LLMs can translate directly into physical risks when these models guide a robot's actions [14]. Consequently, contemporary research also evaluates the efficacy of technical safeguards, such as constraining output tokens [76] or modifying robot behavior [73], to mitigate these multifaceted risks.

- ***Cognitive Load and Workload.*** LLMs introduce a duality to workload assessment in HRI, with the potential to both reduce and create new cognitive demands. Traditional tools like the NASA-TLX questionnaire are still widely used to measure perceived cognitive load [81, 94]. On one hand, LLMs can significantly lower **communication effort** by making instructions more intuitive [42] and enable **cognitive offloading** by taking over tasks for the user [54, 83]. On the other hand, they can introduce **new complexities**, particularly in coordinating multi-agent systems [32] or in safety-critical domains where supervised autonomy must be carefully managed to avoid increasing stress during error recovery. Importantly, users' pre-existing expectations about LLM capabilities can also shape workload perceptions independently of the actual interaction, highlighting a key methodological consideration [124].

7 Applications

Through our analysis, we identified eight principal application domains where LLMs were incorporated into HRI. We classified the existing works into the following high-level clusters: 1) social and conversational systems, 2) healthcare and wellbeing, 3) domestic and everyday use, 4) public spaces service, 5)industrial manufacturing, 6) AR/VR-enabled interactions, 7) teaching and education, 8) other.

Beyond grouping works by their application domains, we further refined each category based on the specific capabilities that LLMs contribute to these scenarios—such as enhanced perception and contextual understanding, generative and agentic interaction capabilities, and iterative optimization and alignment mechanisms. This allowed us to highlight not only where LLMs are used, but how they reshape the functional roles of robots in these contexts. Figure 12 summarizes these categories, including the associated papers and the LLM-enabled sub-capabilities that characterize each domain. The identified domains demonstrate the primary ways LLM integration advances HRI research by addressing longstanding challenges and enabling more adaptive robotic systems.

8 Key Design Considerations and Challenges

Building upon the systematic synthesis of existing literature, this section integrates broader academic perspectives to address **RQ4**. Specifically, we distill eleven design considerations and challenges identified across our corpus. Building on Sense-Interaction-Alignment framework in **Section 4**, these considerations are categorized similarly into three aspects: (1) sense—understanding and perception (challenges 1-4), (2) interaction—action and agency (challenges 5-8), and (3) alignment-adaptation and repair (challenges 9-11).

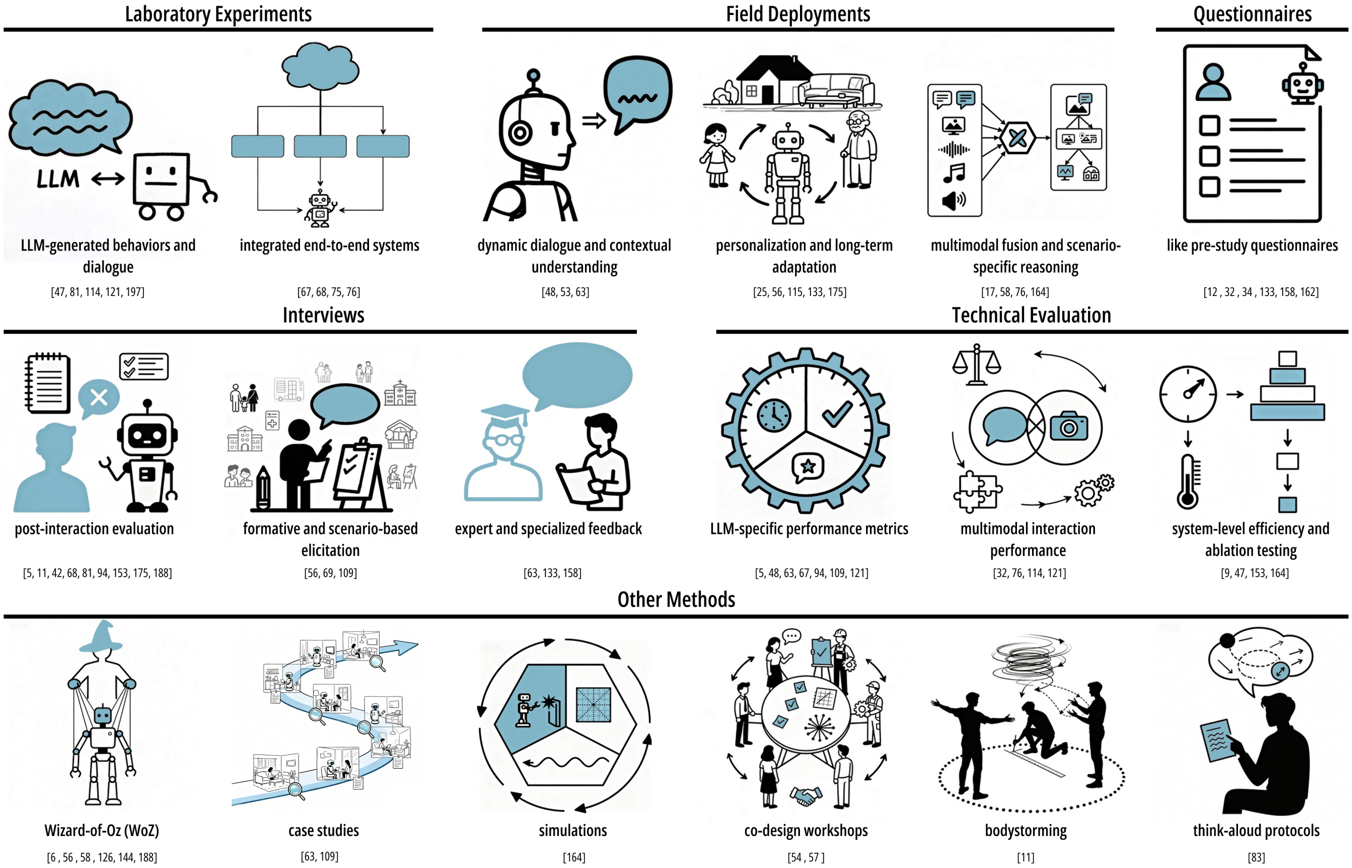


Figure 10: Study methods used in HRI.

- ***Challenge-1. Reliability of LLM-driven Understanding:*** The primary challenge in LLM-driven HRI is the inherent unreliability of robotic understanding, which spans from technical performance to high-level reasoning. First, foundational limitations such as latency, inconsistency, and unpredictability [69, 87, 164] fundamentally undermine the real-time nature of robot perception. Beyond these surface-level issues, a deeper layer of the challenge resides in the fragility of high-level cognitive reasoning, particularly in understanding human social cues. For instance, studies have identified significant failures in LLMs’ ability to interpret humor [48] or perform spatial and quantitative reasoning in embodied contexts [54]. Second, while emerging strategies have attempted to mitigate these shortcomings through grounded prompting, multi-role validation, or human-in-the-loop corrections [27, 65, 68], these interventions often introduce new layers of complexity and opacity. The ultimate challenge remains achieving a level of reliable, interpretable understanding without constant external supervision.

- ***Challenge-2. Multimodal Perception of Emotional Intelligence:*** While the integration of LLMs has enabled robots to respond fluently and enrich interactions through prosody, intonation, and onomatopoeic expressions such as “oh,” “wow,” or “haha” [47, 162], achieving genuine emotional intelligence remains a multi-layered challenge. First, despite advances in empathy calibration and contextual sensitivity in controlled scenarios [34, 67], robots struggle with the instability of multimodal affective signals. In heterogeneous and evolving contexts, signals like hesitation, stress, or collective affect remain inherently noisy and ambiguous [24, 103], making it difficult for robots to perceive emotional nuances consistently beyond surface-level linguistic alignment. Second, a gap persists between generating affective responses and possessing a functional Theory of Mind. Even with hybrid architectures for socio-cognitive grounding [32, 157], most frameworks fail to maintain stability in longitudinal settings, often creating an illusion of understanding rather than genuine intent clarification **(Section 4.1.2)**. Finally, this perceptual gap creates a turn-taking bottleneck [142]. Current systems cannot yet replicate the subtle non-verbal cues (like gaze, fillers, and micropauses) used to negotiate conversational flow. Consequently, bridging fluent generative output with robust multimodal perception remains a critical hurdle for social HRI.

- ***Challenge-3. Multimodal information sensing and alignment:*** While LLM-driven robots rely on integrating various modalities for context-aware interaction (sensing), the core challenge lies in achieving semantic alignment across disparate data streams (e.g., the resolution of modality conflicts). For instance, a robot may face a disconnect when a user’s verbal input (e.g., a joke) contradicts

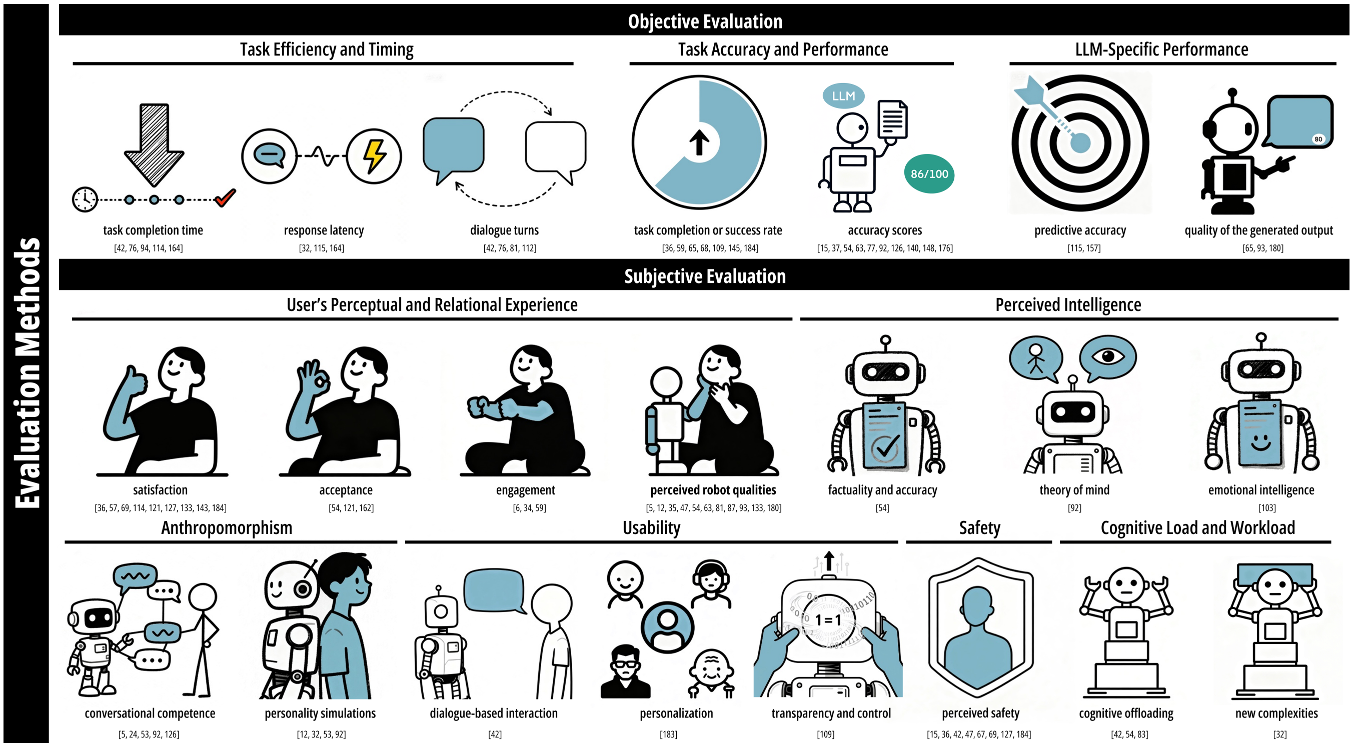
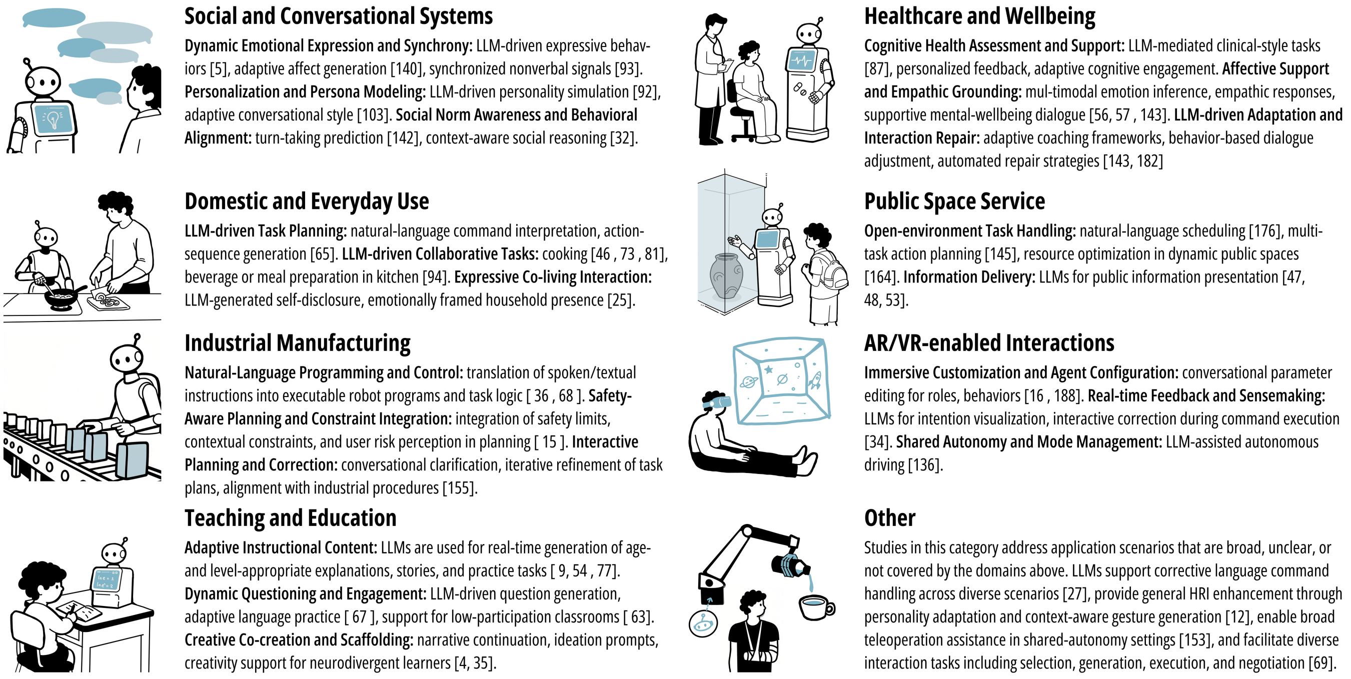


Figure 11: Evaluation methods used in HRI.



Social and Conversational Systems

Dynamic Emotional Expression and Synchrony: LLM-driven expressive behaviors [5], adaptive affect generation [140], synchronized nonverbal signals [93]. **Personalization and Persona Modeling:** LLM-driven personality simulation [92], adaptive conversational style [103]. **Social Norm Awareness and Behavioral Alignment:** turn-taking prediction [142], context-aware social reasoning [32].

Domestic and Everyday Use

LLM-driven Task Planning: natural-language command interpretation, action-sequence generation [65]. **LLM-driven Collaborative Tasks:** cooking [46 , 73 , 81], beverage or meal preparation in kitchen [94]. **Expressive Co-living Interaction:** LLM-generated self-disclosure, emotionally framed household presence [25].

Industrial Manufacturing

Natural-Language Programming and Control: translation of spoken/textual instructions into executable robot programs and task logic [36 , 68]. **Safety-Aware Planning and Constraint Integration:** integration of safety limits, contextual constraints, and user risk perception in planning [15]. **Interactive Planning and Correction:** conversational clarification, iterative refinement of task plans, alignment with industrial procedures [155].

Teaching and Education

Adaptive Instructional Content: LLMs are used for real-time generation of age- and level-appropriate explanations, stories, and practice tasks [9, 54 , 77]. **Dynamic Questioning and Engagement:** LLM-driven question generation, adaptive language practice [67], support for low-participation classrooms [63]. **Creative Co-creation and Scaffolding:** narrative continuation, ideation prompts, creativity support for neurodivergent learners [4, 35].

Healthcare and Wellbeing

Cognitive Health Assessment and Support: LLM-mediated clinical-style tasks [87], personalized feedback, adaptive cognitive engagement. **Affective Support and Empathic Grounding:** mul-timodal emotion inference, empathic responses, supportive mental-wellbeing dialogue [56, 57 , 143]. **LLM-driven Adaptation and Interaction Repair:** adaptive coaching frameworks, behavior-based dialogue adjustment, automated repair strategies [143, 182]

Public Space Service

Open-environment Task Handling: natural-language scheduling [176], multi-task action planning [145], resource optimization in dynamic public spaces [164]. **Information Delivery:** LLMs for public information presentation [47, 48, 53].

AR/VR-enabled Interactions

Immersive Customization and Agent Configuration: conversational parameter editing for roles, behaviors [16 , 188]. **Real-time Feedback and Sensemaking:** LLMs for intention visualization, interactive correction during command execution [34]. **Shared Autonomy and Mode Management:** LLM-assisted autonomous driving [136].

Other

Studies in this category address application scenarios that are broad, unclear, or not covered by the domains above. LLMs support corrective language command handling across diverse scenarios [27], provide general HRI enhancement through personality adaptation and context-aware gesture generation [12], enable broad teleoperation assistance in shared-autonomy settings [153], and facilitate diverse interaction tasks including selection, generation, execution, and negotiation [69].

Figure 12: Usage scenarios and sub-categorization by LLM capabilities.

their physical cues (e.g., an angry facial expression or aggressive posture) [34]. Current LLM-driven architectures often lack the nuanced arbitration logic to weigh these conflicting signals, leading to cascading errors in social interpretation that compromise interactional safety. Furthermore, multimodal alignment requires making the sensing process itself more legible and transparent. Prior work suggests that simple alignment of timestamps is insufficient; instead, the system must communicate its internal interpretation of fused data to the user. As demonstrated by MARCER [59], while transparent feedback via hybrid modalities can mitigate misalignment, the failure to resolve subtle social cue discrepancies often exacerbates user frustration or perceived unfairness. The challenge, therefore, remains in developing robust "cross-modal reasoning" that can dynamically prioritize modalities based on the situational context, rather than merely aggregating them into a singular text-based prompt for the LLM.

- ***Challenge-4. Equitable Engagement in Multi-User Scenarios:*** As LLM-driven robots are increasingly deployed in environments involving multiple users, such as cafés, classrooms, and domestic settings, equitable engagement becomes a critical yet underexplored challenge [49]. To elaborate, robots must not only balance diverse preferences, but also perceive and interpret complex social dynamics among co-present users to avoid conflict and exclusion [109]. Prior work has shown that multi-user settings substantially complicate social sensing; for instance, Skantze and Irfan indicated that multiple users makes it more challenging for robots to determine if the users are addressing the robot or each other [142]. Ethical risks arise when one user attempts to elicit actions that could harm others, such as simulating emergencies in VR studies [144]. Moreover, implicit biases from human-human interactions can carry over into human-robot interactions when robots lack sufficient social awareness to detect and compensate for unequal engagement patterns, thereby exacerbating inequities in multi-user contexts. To mitigate these, robots could integrate seamlessly into the social and informational frameworks of their environment rather than operate as isolated agents, ensuring responsible coordination, fairness, and safety across users [162].

- ***Challenge-5. Morphology-Aligned Social Intelligence:*** Morphology constrains the range of behavioral expressions a robot can produce, thereby shaping how social agency is physically expressed during interaction. LLMs enable robots to interpret semantic meaning and engage in flexible dialogue, which in turn raises expectations for the robot's morphology (e.g., head, arms) to express corresponding social and emotional capabilities [70]. For example, LLMs can be used to generate or refine expressive robot behaviors such as nodding [153], as well as to produce richly articulated motion sequences [81, 93] **(Section 4.2.1)**. However, this enhancement is not a linear progression toward unconditionally anthropomorphizing robots [25]. Achieving fully human-like expressiveness remains technically challenging and ethically fraught [114]. Mismatches between highly fluent linguistic output and comparatively rudimentary physical behavior can produce expectation gaps [49, 53]; LLM hallucinations may trigger trust ruptures [76, 170]; and overly anthropomorphic presentations may heighten user discomfort like uncanny valley [127]. Consequently, we suggest future HRI design carefully balance LLM-driven social agency with the morphology limitations through deliberate design choices, such as prioritizing physical compliance over anthropomorphic realism [15] or incorporating explicit physical cues [162, 188], to mitigate risks associated with over-humanization.

- ***Challenge-6. Balancing Autonomy and Human Oversight:*** LLMs substantially expand robots' autonomy by enhancing natural language control [36, 65, 107], dynamic code generation [15, 68], and common sense reasoning [93, 178]. However, this shift raises challenges for balancing LLM-driven agency with human control. First, high-granularity end-user programming can cause robots to be perceived as tools rather than autonomous partners, reducing the sense of intelligence and social presence [183]. Second, while LLM-driven automation may increase task efficiency, it can shift focus toward individual outcomes and inadvertently suppress opportunities for meaningful communication with other people [136]. Third, socially proactive behaviors (e.g., small talk and emotional responses) may impose unintended interactional obligations on users [114]. At the same time, some domains (e.g., assistive and manipulation tasks) benefit from shared autonomy, where users can provide natural-language online corrections to refine robot behavior [27] **(Section 4.2.2)**. This highlights the need for hybrid autonomy designs that balance LLM-driven initiative with task efficiency [136]. Moving forward, autonomy could be structured around shared-control or Human-in-the-Loop frameworks, ensuring that LLM-driven autonomy remains aligned with user intent even in complex or high-stakes environments [5, 81, 87].

- ***Challenge-7. Balancing Trust and Overtrust:*** In LLM-driven HRI, trust mediates the delegation of authority to robots, yet LLMs often exacerbate overtrust and overreliance [11, 54, 63, 75, 81]. First, trust is induced multidimensionally; as synthesized in **Section 4.2.1**, factors like personalization [12, 103], social norm awareness [32, 34], and human-likeness [5, 25] lower trust thresholds. Beyond factual accuracy, this process requires contextual grounding and expectation management [170]; however, anthropomorphic cues often trigger social interpretations that exceed actual system reliability. Second, technical opacity hinders trust calibration. While objective metrics like the Attention Arbitration Ratio can predict trust levels [46], technical fragilities (like hallucinations and inconsistent reasoning) foster uncalibrated overtrust [42, 83, 121]. This is critical when perceived confidence masks underlying uncertainty [75, 136], causing difficult-to-mitigate trust ruptures. Finally, trust repair remains asymmetric. Despite multi-level strategies involving apologies or explanations [6] (**Section 4.3.2**), effectively recalibrating lost confidence is still an open question.

- ***Challenge-8. Safeguarding Privacy and Mitigating Safety Risks:*** In LLM-driven HRI, privacy and safety transcend traditional cybersecurity as robots act as embodied observers in private spaces. First, LLMs enable semantic privacy risks: rather than just capturing raw data, robots perform continuous multimodal reasoning to infer user habits and social dynamics [133, 159]. This "semantic surveillance" creates intrusive risks of accidental data exposure based on deep contextual understanding. Second, LLMs as decision engines introduce unpredictable physical risks. Model hallucinations or opacity can lead to hazardous actions, such as misinterpreting intent in high-stakes tasks [87, 136]. While strategies like prompt deletion or age-specific guidelines exist [4, 49], these reactive measures struggle with the inherent unpredictability of an agent physicalizing its reasoning. Finally, the contextual variability of safety norms across cultures complicates the design of universal frameworks. Consequently, safeguarding LLM-driven HRI requires moving beyond encryption toward frameworks addressing the unique vulnerabilities of embodied, context-aware interaction.

- ***Challenge-9. Stabilizing Personalization and Social Alignment:*** Personalization has been widely praised for sustaining user enjoyment [67, 143, 162] **(Section 4.3.1)**. However, personalization can also create unintended emotional dependence in vulnerable users, where the sudden withdrawal of robots leads to significant harm [56]. Empathic calibration strategies, though powerful in the short term, risk losing their novelty and eroding trust over time **(Section 4.2)**. For instance, affective styles are central to sustaining comfort and naturalness in short-term interactions [68, 114, 164], yet the inconsistency of LLMs (e.g., humor understanding) challenges the reliability of such strategies [48]. Similarly, the unpredictability of LLM outputs raises risks for repair strategies: while transparency and empathy are expected in many domains [6, 87, 133], overly elaborate or irrelevant repairs can disrupt conversational flow [11, 114]. Moreover, some users may actually prefer robotic companionship precisely because it bypasses the unpredictability and emotional complexity of human relationships [162]. If these preferences become widespread, how designers should balance personalization-driven alignment with long-term user well-being remains an open question.

- ***Challenge-10. Sustaining Long-Term Engagement:*** Effective integration of LLM-driven robots into daily routines hinges on sustaining engagement beyond the initial novelty effect [58]. The first challenge lies in evolving personalization; while LLMs support varied interactions in healthcare and domestic settings [76, 159], existing systems often fail to transform short-term memory into a stable, growing personality. Current behavioral modeling methods **(Section 4.3.1)** struggle to keep pace with shifting human expectations, leading to "stale" interactions as the honeymoon phase fades. Secondly, a significant methodological gap persists in longitudinal validation. Despite emerging field studies in real-world environments **(Section 6.1.2)**, most empirical evidence remains limited to "snapshot" observations. As noted in the SET-PAiREd [54], brief deployments fail to capture how roles and trust dynamics evolve as LLM integrates into the domestic fabric. This lack of sustained data prevents a systematic understanding of adaptation fatigue over time. Finally, the transition from tools to persistent companions raises profound ethical concerns regarding emotional dependency and value drift. Thus, the challenge lies in designing frameworks that co-evolve with users while safeguarding social well-being and ethical boundaries in longitudinal settings.

- ***Challenge-11. Proactive Repair for Diverse Contexts:*** Repair presents a critical design space for LLM-driven HRI. Robots today can already leverage dialogue history to provide explanations, apologies, and repair strategies [6] **(Section 4.3.2)**. However, repair strategies cannot remain reactive templates; they must become proactive and adaptive, evolving with longitudinal expectations [4, 11, 42, 75, 107, 114, 121]. In-context learning provides early evidence for this shift, demonstrating that robots can anticipate errors, distinguish genuine interruptions from noise, propose grounded alternatives, and transparently explain their limitations [81, 109, 121]. This reconfiguration moves repair from a corrective action into a continuous socio-emotional negotiation that preserves trust and transparency across long-term engagement [11, 24, 42, 121]. However, there are still limited investigations in how such adaptive repair mechanisms can be calibrated to different domains, from casual conversation to high-stakes contexts.

9 Discussion and Limitations

Several limitations arise from our corpus construction and assessment process. First, although our search terms were iteratively refined to cover major HRI and robotics venues, they could not fully capture the breadth of emerging work across interdisciplinary domains. This may have resulted in the under-representation of studies from psychology, sociology, and communication research that examine LLM-mediated social interaction and human–machine boundaries. Further, our inclusion criteria included studies with both explicit (e.g., LLMs as system components) and implicit (e.g., LLMs as tools for experimental design) applications, which may introduce heterogeneity to the data collection. For example, Rosén et al. [124] used GPT-3 to replace Wizard-of-Oz methods to avoid introducing human actor expectations, while Grassi et al. [49] employed LLMs to increase response variability. Compared with explicit use, such implicit uses were not central to the studies and often lacked rigorous evaluation, limiting cross-study comparability. Moreover, the assessment of conceptually ambiguous and borderline studies inevitably reflects our own interpretive stance, and other researchers might reasonably draw different boundaries. Future reviews may benefit from broader interdisciplinary coverage and differentiated treatment of implicit versus explicit LLM integration.

In addition to the scoping of research themes, our selection of publication types also warrants discussion. We excluded Late-Breaking Reports (LBRs) [6] to ensure our dataset primarily reflects rigorously validated, methodologically comprehensive research. While we acknowledge LBRs often showcase cutting-edge LLM applications in HRI, focusing on full archival papers prioritizes synthesis reliability over preliminary findings. This trade-off ensures a robust foundation for our analysis. Future reviews could specifically synthesize these emerging LBRs to capture the field's rapid, real-time evolution. It has been truly inspiring to witness and synthesize the rapid evolution of this flourishing field through our work. We hope this review contributes to the collective endeavor of building more intelligent, human-centric robotic systems.

10 Conclusion

In this paper, we present our systematic review and taxonomy of LLM-driven HRI research, synthesizing existing research approaches, interaction designs, and evaluation strategies across identified studies. Our goal is to provide a common ground for researchers to understand how LLMs are shaping HRI systems. Building on this synthesis, we introduce the Sense–Interaction–Alignment framework to consolidate how LLMs enable new embodied capabilities, including enhanced contextual sensing, generative and socially grounded interaction, and continuous human-aligned adaptation across scenarios. In addition, to further stimulate future research at the intersection of LLMs and HRI, we discuss key design considerations and emerging challenges, such as multimodal grounding,

[6] Late-Breaking Reports are a publication category in the ACM/IEEE International Conference on Human–Robot Interaction (HRI).

morphology-aligned social intelligence, and sustaining long-term alignment in real-world settings. We hope our review, taxonomy, and identified research directions will guide and inspire future work in LLM-driven HRI.

11 Disclosure about Use of LLM

Portions of this paper were refined using GPT and DeepL for clarity and grammatical accuracy; all content has been thoroughly proofread by the authors.

Acknowledgments

We gratefully acknowledge the support of the Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things Grant (No. 2023B1212010007), the 111 Center Grant (No. D25008), and Natural Science Foundation of China (No. 62402271). This work was completed during the internships of Yufeng Wang, Anastasia Nikolova, and Yuxuan Wang in the Human-Centered Intelligence+ Summer Research Program, and we sincerely thank the program and support team.

References

[1] Munir Ahmad, Majid Habib Khan, Assia Bouabdallah, Nora Zemoura, and Muhammad Imran. 2024. Exploring Human-Robot Interaction and Collaboration for Real-World Applications. *IEEE-SEM* 12, 2 (2024), 52–60.

[2] Arash Ajoudani, Andrea Maria Zanchettin, Serena Ivaldi, Alin Albu-Schäffer, Kazuhiro Kosuge, and Oussama Khatib. 2018. Progress and Prospects of the Human–Robot Collaboration. *Autonomous Robots* 42, 5 (June 2018), 957–975. doi:10.1007/s10514-017-9677-2

[3] R. Alami, A. Albu-Schaeffer, A. Bicchi, R. Bischoff, R. Chatila, A. De Luca, A. De Santis, G. Giralt, J. Guiochet, G. Hirzinger, F. Ingrand, V. Lippiello, R. Mattone, D. Powell, S. Sen, B. Siciliano, G. Tonietti, and L. Villani. 2006. Safe and dependable physical human-robot interaction in anthropic domains: State of the art and challenges. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1–16. doi:10.1109/IROS.2006.6936985

[4] Safinah Ali, Ayat Abodayeh, Zahra Dhuliawala, Cynthia Breazeal, and Hae Won Park. 2025. Towards Inclusive Co-creative Child-robot Interaction: Can Social Robots Support Neurodivergent Children's Creativity?. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 321–330. doi:10.1109/HRI61500.2025.10974006

[5] Victor Nikhil Antony, Maia Stiber, and Chien-Ming Huang. 2025. Xpress: A System For Dynamic, Context-Aware Robot Facial Expressions using Language Models. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 958–967. doi:10.1109/HRI61500.2025.10974045

[6] Mehdi Arjmand, Farnaz Nouraei, Ian Steenstra, and Timothy Bickmore. 2024. Empathic Grounding: Explorations using Multimodal Interaction and Large Language Models with Conversational Agents. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents* (GLASGOW, United Kingdom) *(IVA '24)*. Association for Computing Machinery, New York, NY, USA, Article 6, 10 pages. doi:10.1145/3652988.3673949

[7] IEEE Standards Association. 2015. IEEE Standard Ontologies for Robotics and Automation. IEEE Std 1872-2015, 60 pages. doi:10.1109/IEEESTD.2015.7084073

[8] Jesse Atuhurra. 2024. Leveraging Large Language Models in Human-Robot Interaction: A Critical Analysis of Potential and Pitfalls. arXiv:2405.00693 [cs.RO] https://arxiv.org/abs/2405.00693

[9] Agnes Axelsson and Gabriel Skantze. 2023. Do You Follow? A Fully Automated System for Adaptive Robot Presenters. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 102–111. doi:10.1145/3568162.3576958

[10] Minja Axelsson, Nikhil Churamani, Atahan Çaldır, and Hatice Gunes. 2025. Participant Perceptions of a Robotic Coach Conducting Positive Psychology Exercises: A Qualitative Analysis. *J. Hum.-Robot Interact.* 14, 2, Article 36 (March 2025), 27 pages. doi:10.1145/3711937

[11] Minja Axelsson, Micol Spitale, and Hatice Gunes. 2024. "Oh, Sorry, I Think I Interrupted You": Designing Repair Strategies for Robotic Longitudinal Wellbeing Coaching. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '24)*. Association for Computing Machinery, New York, NY, USA, 13–22. doi:10.1145/3610977.3634948

[12] Tahsin Tariq Banna, Sejuti Rahman, and Mohammad Tareq. 2025. Beyond Words: Integrating Personality Traits and Context-Driven Gestures in Human-Robot Interactions. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems* (Detroit, MI, USA) *(AAMAS '25)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 242–251.

[13] Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. 2020. *Human-Robot Interaction: An Introduction*. Cambridge University Press.

[14] Abdelrhman Bassiouny, Ahmed H. Elsayed, Zoe Falomir, and Angel P. del Pobil. 2025. UJI-Butler: A Symbolic/Non-symbolic Robotic System That Learns Through Multi-modal Interaction. *International Journal of Social Robotics* 17, 12 (Dec. 2025), 2883–2903. doi:10.1007/s12369-025-01234-5

[15] Brieuc Bastin, Shoichi Hasegawa, Jorge Solis, Renaud Ronsse, Benoit Macq, Lotfi El Hafi, Gustavo Alfonso Garcia Ricardez, and Tadahiro Taniguchi. 2025. GPTAlly: A Safety-Oriented System for Human-Robot Collaboration Based on Foundation Models. In *2025 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 878–884. doi:10.1109/SII59315.2025.10870936

[16] Andrea Bellucci, Giulio Jacucci, Kien Duong Trung, Pritom Kumar Das, Sergei Viktorovich Smirnov, Imtiaj Ahmed, and Jean-Luc Lugrin. 2025. Immersive Tailoring of Embodied Agents Using Large Language Models. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, Saint Malo, France, 392–400. doi:10.1109/VR59515.2025.00063

[17] Antonio Blanco, Gerardo Pérez, Alicia Condón, Trinidad Rodríguez, and Pedro Núñez. 2024. AI-Enhanced Social Robots for Older Adults Care: Evaluating the Efficacy of ChatGPT-Powered Storytelling in the EBO Platform. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, 2109–2116. doi:10.1109/RO-MAN60168.2024.10731292

[18] Chandhawat Boonyard, Christophe Jouffrais, Jessica R. Cauchard, and Anke M. Brock. 2025. Firefighting with Drone Assistance: User Needs and Design Considerations for Thailand. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 432, 18 pages. doi:10.1145/3706598.3714172

[19] Giovanni Boschetti, Maurizio Faccio, and Irene Granata. 2023. Human-Centered Design for Productivity and Safety in Collaborative Robots Cells: A New Methodological Approach. *Electronics* 12, 167 (2023). doi:10.3390/electronics12010167

[20] Anya Bouzida, Alyssa Kubota, Dagoberto Cruz-Sandoval, Elizabeth W. Twamley, and Laurel D. Riek. 2024. CARMEN: A Cognitively Assistive Robot for Personalized Neurorehabilitation at Home. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '24)*. Association for Computing Machinery, New York, NY, USA, 55–64. doi:10.1145/3610977.3634971

[21] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[22] Fanjun Bu, Alexandra W.D. Bremers, Mark Colley, and Wendy Ju. 2024. Field Notes on Deploying Research Robots in Public Spaces. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 162, 6 pages. doi:10.1145/3613905.3651044

[23] Giulio Campagna and Matthias Rehm. 2025. A Systematic Review of Trust Assessments in Human–Robot Interaction. *J. Hum.-Robot Interact.* 14, 2, Article 30 (Jan. 2025), 35 pages. doi:10.1145/3706123

[24] Hyungjun Cho, Jiyeon Lee, Bonhee Ku, Yunwoo Jeong, Shakhnozakhon Yadgarova, and Tek-Jin Nam. 2023. ARECA: A Design Speculation on Everyday Products Having Minds. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 31–44. doi:10.1145/3563657.3596002

[25] Hyungjun Cho and Tek-Jin Nam. 2025. Living Alongside Areca: Exploring Human Experiences with Things Expressing Thoughts and Emotions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 434, 16 pages. https://doi.org/10.1145/3706598.3713228

[26] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. doi:10.1177/001316446002000104

[27] Yuchen Cui, Siddharth Karamcheti, Raj Palleti, Nidhya Shivakumar, Percy Liang, and Dorsa Sadigh. 2023. No, to the Right: Online Language Corrections for Robotic Manipulation via Shared Autonomy. In *Proceedings of the 2023 ACM/IEEE*

International Conference on Human-Robot Interaction (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 93–101. doi:10.1145/3568162.3578623

[28] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies: why and how. In *Proceedings of the 1st International Conference on Intelligent User Interfaces* (Orlando, Florida, USA) *(IUI '93)*. Association for Computing Machinery, New York, NY, USA, 193–200. doi:10.1145/169891.169968

[29] Kerstin Dautenhahn. 2007. Methodology & Themes of Human-Robot Interaction: A Growing Research Field. *International Journal of Advanced Robotic Systems* 4, 1 (2007), 15. doi:10.5772/5702

[30] Kerstin Dautenhahn. 2007. Socially Intelligent Robots: Dimensions of Human–Robot Interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences* 362, 1480 (Feb. 2007), 679. doi:10.1098/rstb.2006.2004

[31] Alessandro De Luca and Fabrizio Flacco. 2012. Integrated control for pHRI: Collision avoidance, detection, reaction and collaboration. In *2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*. IEEE, 288–295. doi:10.1109/BioRob.2012.6290917

[32] Davide Dell'Anna and Anahita Jamshidnejad. 2024. SONAR: An Adaptive Control Architecture for Social Norm Aware Robots. *International Journal of Social Robotics* 16, 9 (Oct. 2024), 1969–2000. doi:10.1007/s12369-024-01172-8

[33] Wen Duan, Christopher Flathmann, Nathan McNeese, Matthew J Scalia, Ruihao Zhang, Jamie Gorman, Guo Freeman, Shiwen Zhou, Allyson Ivy Hauptman, and Xiaoyun Yin. 2025. Trusting Autonomous Teammates in Human-AI Teams - A Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1102, 23 pages. doi:10.1145/3706598.3713527

[34] Morad Elfleet and Mathieu Chollet. 2024. Investigating the Impact of Multimodal Feedback on User-Perceived Latency and Immersion with LLM-Powered Embodied Conversational Agents in Virtual Reality. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents* (GLASGOW, United Kingdom) *(IVA '24)*. Association for Computing Machinery, New York, NY, USA, Article 12, 9 pages. doi:10.1145/3652988.3673965

[35] Maha Elgarf, Sahba Zojaji, Gabriel Skantze, and Christopher Peters. 2022. CreativeBot: a Creative Storyteller robot to stimulate creativity in children. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 540–548. doi:10.1145/3536221.3556578

[36] Muhammad Umar Farooq, Geon Kang, Jiwon Seo, Jungchan Bae, Seoyeon Kang, and Young Jae Jang. 2024. DAIM-HRI: A new Human-Robot Integration Technology for Industries. In *2024 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. IEEE, 7–12. doi:10.1109/ARSO60199.2024.10557811

[37] Lorenzo Ferrini, Antonio Andriella, Raquel Ros, and Séverin Lemaignan. 2025. From Percepts to Semantics: A Multi-modal Saliency Map to Support Social Robots' Attention. *J. Hum.-Robot Interact.* 14, 4, Article 71 (July 2025), 19 pages. doi:10.1145/3737891

[38] Debora Firmino de Souza, Sonia Sousa, Kadri Kristjuhan-Ling, Olga Dunajeva, Mare Roosileht, Avar Pentel, Mati Mõttus, Mustafa Can Özdemir, and Žanna Gratšjova. 2025. Trust and Trustworthiness from Human-Centered Perspective in Human–Robot Interaction (HRI)—A Systematic Literature Review. *Electronics* 14, 8 (Jan. 2025), 1557. doi:10.3390/electronics14081557

[39] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, Brian Ichter, Danny Driess, Jiajun Wu, Cewu Lu, and Mac Schwager. 2025. Foundation Models in Robotics: Applications, Challenges, and the Future. *The International Journal of Robotics Research* 44, 5 (April 2025), 701–739. doi:10.1177/02783649241281508

[40] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds Mach.* 30, 4 (Dec. 2020), 681–694. doi:10.1007/s11023-020-09548-1

[41] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A Survey of Socially Interactive Robots. *Robotics and Autonomous Systems* 42, 3-4 (March 2003), 143–166. doi:10.1016/S0921-8890(02)00372-X

[42] Yate Ge, Meiying Li, Xipeng Huang, Yuanda Hu, Qi Wang, Xiaohua Sun, and Weiwei Guo. 2025. GenComUI: Exploring Generative Visual Aids as Medium to Support Task-Oriented Human-Robot Communication. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 433, 21 pages. doi:10.1145/3706598.3714238

[43] Gemini Robotics Team et al. 2025. Gemini Robotics: Bringing AI into the Physical World. arXiv:2503.20020 [cs.RO] https://arxiv.org/abs/2503.20020

[44] Christos Gkournelos, Christos Konstantinou, and Sotiris Makris. 2024. An LLM-based approach for enabling seamless Human-Robot collaboration in assembly. *CIRP Annals* 73, 1 (2024), 9–12. doi:10.1016/j.cirp.2024.04.002

[45] Michael A. Goodrich and Alan C. Schultz. 2007. Human-Robot Interaction: A Survey. *Foundations and Trends® in Human-Computer Interaction* 1, 3 (2007), 203–275. doi:10.1561/1100000005

[46] Cedric Goubard and Yiannis Demiris. 2025. Cognitive Modelling of Visual Attention Captures Trust Dynamics in Human–Robot Collaboration. *J. Hum.-Robot Interact.* 14, 4, Article 67 (July 2025), 21 pages. doi:10.1145/3732795

[47] Lucrezia Grassi, Zhouyang Hong, Carmine Tommaso Recchiuto, and Antonio Sgorbissa. 2024. Grounding Conversational Robots on Vision Through Dense Captioning and Large Language Models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5492–5498. doi:10.1109/ICRA57147.2024.10611232

[48] Lucrezia Grassi, Carmine Tommaso Recchiuto, and Antonio Sgorbissa. 2024. Enhancing LLM-Based Human-Robot Interaction with Nuances for Diversity Awareness. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, 2287–2294. doi:10.1109/RO-MAN60168.2024.10731381

[49] Lucrezia Grassi, Carmine Tommaso Recchiuto, and Antonio Sgorbissa. 2025. Strategies for Controlling the Conversation Dynamics in Multi-Party Human-Robot Interaction. *International Journal of Social Robotics* 17, 8 (Aug. 2025), 1517–1539. doi:10.1007/s12369-025-01298-3

[50] Johanna Gunawan, Sarah Elizabeth Gillespie, David Choffnes, Woodrow Hartzog, and Christo Wilson. 2025. Promises, Promises: Understanding Claims Made in Social Robot Consumer Experiences. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 601, 22 pages. doi:10.1145/3706598.3713471

[51] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large Language Model Based Multi-agents: A Survey of Progress and Challenges. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Jeju, South Korea, 8048–8057. doi:10.24963/ijcai.2024/890

[52] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-Trained Models: Past, Present and Future. arXiv:2106.07139 [cs.AI] https://arxiv.org/abs/2106.07139

[53] Damith Herath, Janie Busby Grant, Adrian Rodriguez, and Jenny L. Davis. 2025. First Impressions of a Humanoid Social Robot with Natural Language Capabilities. *Scientific Reports* 15, 1 (June 2025), 19715. doi:10.1038/s41598-025-04274-z

[54] Hui-Ru Ho, Nitigya Kargeti, Ziqi Liu, and Bilge Mutlu. 2025. SET-PAiREd: Designing for Parental Involvement in Learning with an AI-Assisted Educational Robot. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1040, 20 pages. doi:10.1145/3706598.3713330

[55] Damian Hostettler, Simon Mayer, Jan Liam Albert, Kay Erik Jenss, and Christian Hildebrand. 2025. Real-Time Adaptive Industrial Robots: Improving Safety And Comfort In Human-Robot Collaboration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 908, 16 pages. doi:10.1145/3706598.3713889

[56] Long-Jing Hsu, Janice Bays, Manasi Swaminathan, Weslie Khoo, Hiroki Sato, Kyrie Jig Amon, Sathvika Dobbala, Min Min Thant, Alex Foster, Kate Tsui, Philip B. Stafford, David Crandall, and Selma Sabanovic. 2025. Research as Care: A Reflection on Incorporating the Ethics of Care in Design Research with People Living with Dementia. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference* (Funchal, Portugal) *(DIS '25)*. Association for Computing Machinery, New York, NY, USA, 3013–3027. doi:10.1145/3715336.3735678

[57] Long-Jing Hsu, Manasi Swaminathan, Weslie Khoo, Kyrie Jig Amon, Hiroki Sato, Sathvika Dobbala, Kate Tsui, David Crandall, and Selma Sabanovic. 2025. Bittersweet Snapshots of Life: Designing to Address Complex Emotions in a Reminiscence Interaction between Older Adults and a Robot. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1030, 18 pages. doi:10.1145/3706598.3714256

[58] Yaxin Hu, Anjun Zhu, Catalina L. Toma, and Bilge Mutlu. 2025. Designing Telepresence Robots to Support Place Attachment. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 252–261. doi:10.1109/HRI61500.2025.10974101

[59] Bryce Ikeda, Maitrey Gramopadhye, LillyAnn Nekervis, and Daniel Szafir. 2025. MARCER: Multimodal Augmented Reality for Composing and Executing Robot Tasks. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 529–539. doi:10.1109/HRI61500.2025.10974232

[60] Sara Incao, Carlo Mazzola, Giulia Belgiovine, and Alessandra Sciutti. 2025. *A Roadmap for Embodied and Social Grounding in LLMs*. Frontiers in Artificial Intelligence and Applications, Vol. 397. IOS Press, 43–52. doi:10.3233/FAIA241488

[61] ISO. 2019. 9241–210:2019 Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems. https://www.iso.org/standard/

77520.html

[62] ISO. 2021. 8373:2021 Robotics — Vocabulary. https://www.iso.org/obp/ui/#iso:std:iso:8373:ed-3:v1:en

[63] Shunichiro Ito, Kanae Kochigami, and Takayuki Kanda. 2025. A Robot Dynamically Asking Questions in University Classes. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 839–848. doi:10.1109/HRI61500.2025.10973850

[64] Hyeongyo Jeong, Haechan Lee, Changwon Kim, and Sungtae Shin. 2024. A Survey of Robot Intelligence with Large Language Models. *Applied Sciences* 14, 19 (Jan. 2024), 8868. doi:10.3390/app14198868

[65] Yixiang Jin, Dingzhe Li, Yong A, Jun Shi, Peng Hao, Fuchun Sun, Jianwei Zhang, and Bin Fang. 2024. RobotGPT: Robot Manipulation Learning From ChatGPT. *IEEE Robotics and Automation Letters* 9, 3 (March 2024), 2543–2550. doi:10.1109/LRA.2024.3357432

[66] Céline Jost, Brigitte Le Pévédic, Tony Belpaeme, Cindy Bethel, Dimitrios Chrysostomou, Nigel Crook, Marine Grandgeorge, and Nicole Mirnig (Eds.). 2020. *Human-Robot Interaction: Evaluation Methods and Their Standardization*. Springer Series on Bio- and Neurosystems, Vol. 12. Springer International Publishing, Cham. doi:10.1007/978-3-030-42307-0

[67] Alireza M. Kamelabad, Elin Inoue, and Gabriel Skantze. 2025. Comparing Monolingual and Bilingual Social Robots as Conversational Practice Companions in Language Learning. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 829–838. doi:10.1109/HRI61500.2025.10973901

[68] Ulas Berk Karli, Juo-Tung Chen, Victor Nikhil Antony, and Chien-Ming Huang. 2024. Alchemist: LLM-Aided End-User Development of Robot Applications. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '24)*. Association for Computing Machinery, New York, NY, USA, 361–370. doi:10.1145/3610977.3634969

[69] Callie Y. Kim, Christine P. Lee, and Bilge Mutlu. 2024. Understanding Large-Language Model (LLM)-powered Human-Robot Interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '24)*. Association for Computing Machinery, New York, NY, USA, 371–380. doi:10.1145/3610977.3634966

[70] J. Taery Kim, Morgan Byrd, Jack L Crandell, Bruce N. Walker, Greg Turk, and Sehoon Ha. 2025. Understanding Expectations for a Robotic Guide Dog for Visually Impaired People. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 262–271. doi:10.1109/HRI61500.2025.10974141

[71] Kyungki Kim, John Windle, Melissa Christian, Tom Windle, Erica Ryherd, Pei-Chi Huang, Anthony Robinson, and Reid Chapman. 2024. Framework for Integrating Large Language Models with a Robotic Health Attendant for Adaptive Task Execution in Patient Care. *Applied Sciences* 14, 9922 (2024). doi:10.3390/app14219922

[72] Yeseung Kim, Dohyun Kim, Jieun Choi, Jisang Park, Nayoung Oh, and Daehyung Park. 2024. A Survey on Integration of Large Language Models with Intelligent Robots. *Intelligent Service Robotics* 17, 5 (Sept. 2024), 1091–1107. arXiv:2404.09228 [cs] doi:10.1007/s11370-024-00550-5

[73] Krishna Kodur, Manizheh Zand, Matthew Tognotti, Sean Banerjee, Natasha Kholgade Banerjee, and Maria Kyrarini. 2025. Exploring the Dynamics of Human-Robot Interaction: Robot Error, Sentiment Analysis, and Politeness. *International Journal of Social Robotics* 17, 7 (July 2025), 1221–1234. doi:10.1007/s12369-025-01282-x

[74] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1613, 15 pages.

[75] Dimosthenis Kontogiorgos and Julie Shah. 2025. Questioning the Robot: Using Human Non-verbal Cues to Estimate the Need for Explanations. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 717–728. doi:10.1109/HRI61500.2025.10974079

[76] Yuzhi Lai, Shenghai Yuan, Youssef Nassar, Mingyu Fan, Atmaraaj Gopal, Arihiro Yorita, Naoyuki Kubota, and Matthias Rätsch. 2025. Natural Multimodal Fusion-Based Human–Robot Interaction: Application With Voice and Deictic Posture via Large Language Model. *IEEE Robotics & Automation Magazine* (2025), 2–11. doi:10.1109/MRA.2025.3543957

[77] Ehsan Latif, Ramviyas Parasuraman, and Xiaoming Zhai. 2024. PhysicsAssistant: An LLM-Powered Interactive Learning Robot for Physics Lab Investigations. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, 864–871. doi:10.1109/RO-MAN60168.2024.10731312

[78] Steven Lawrence, Melanie Jouaiti, Jesse Hoey, Chrystopher L. Nehaniv, and Kerstin Dautenhahn. 2025. The Role of Social Norms in Human–Robot Interaction: A Systematic Review. *J. Hum.-Robot Interact.* 14, 3, Article 56 (June 2025), 44 pages. doi:10.1145/3722120

[79] Benedikt Leichtmann, Verena Nitsch, and Martina Mara. 2022. Crisis Ahead? Why Human-Robot Interaction User Studies May Have Replicability Problems and Directions for Improvement. *Frontiers in Robotics and AI* 9 (March 2022). doi:10.3389/frobt.2022.838116

[80] Gregory LeMasurier, Christian Tagliamonte, Jacob Breen, Daniel Maccaline, and Holly A. Yanco. 2024. Templated vs. Generative: Explaining Robot Failures. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, Pasadena, CA, USA, 1346–1353. doi:10.1109/RO-MAN60168.2024.10731331

[81] Jan Leusmann, Anna Belardinelli, Luke Haliburton, Stephan Hasler, Albrecht Schmidt, Sven Mayer, Michael Gienger, and Chao Wang. 2025. Investigating LLM-Driven Curiosity in Human-Robot Interaction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 599, 16 pages. doi:10.1145/3706598.3713923

[82] Chen Li, Xiaochun Zhang, Dimitrios Chrysostomou, and Hongji Yang. 2022. ToD4IR: A Humanised Task-Oriented Dialogue System for Industrial Robots. *IEEE Access* 10 (2022), 91631–91649. doi:10.1109/ACCESS.2022.3202554

[83] Jiannan Li, Maurício Sousa, Karthik Mahadevan, Bryan Wang, Paula Akemi Aoyagui, Nicole Yu, Angela Yang, Ravin Balakrishnan, Anthony Tang, and Tovi Grossman. 2023. Stargazer: An Interactive Camera Robot for Capturing How-To Videos Based on Subtle Instructor Cues. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 800, 16 pages. doi:10.1145/3544548.3580896

[84] Ming Li, Keyu Chen, Ziqian Bi, Ming Liu, Xinyuan Song, Zekun Jiang, Tianyang Wang, Benji Peng, Qian Niu, Junyu Liu, Jinlang Wang, Sen Zhang, Xuanhe Pan, Jiawei Xu, and Pohsun Feng. 2025. Surveying the MLLM Landscape: A Meta-Review of Current Surveys. arXiv:2409.18991 [cs.CL] https://arxiv.org/abs/2409.18991

[85] Peihan Li, Zijian An, Shams Abrar, and Lifeng Zhou. 2025. Large Language Models for Multi-Robot Systems: A Survey. arXiv:2502.03814 [cs.RO] https://arxiv.org/abs/2502.03814

[86] Jonghan Lim, Sujani Patel, Alex Evans, John Pimley, Yifei Li, and Ilya Kovalenko. 2024. Enhancing Human-Robot Collaborative Assembly in Manufacturing Systems Using Large Language Models. In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2581–2587. doi:10.1109/CASE59546.2024.10711843

[87] Maria R. Lima, Amy O'Connell, Feiyang Zhou, Alethea Nagahara, Avni Hulyalkar, Anura Deshpande, Jesse Thomason, Ravi Vaidyanathan, and Maja Matarić. 2025. Promoting Cognitive Health in Elder Care with Large Language Model-Powered Socially Assistive Robots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 317, 22 pages. doi:10.1145/3706598.3713582

[88] Ming-Yi Lin, Ou-Wen Lee, and Chih-Ying Lu. 2024. Embodied AI with Large Language Models: A Survey and New HRI Framework. In *2024 International Conference on Advanced Robotics and Mechatronics (ICARM)*. IEEE, Tokyo, Japan, 978–983. doi:10.1109/ICARM62033.2024.10715872

[89] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2025. Aligning Cyber Space With Physical World: A Comprehensive Survey on Embodied AI. *IEEE/ASME Transactions on Mechatronics* 30, 6 (2025), 7253–7274. doi:10.1109/TMECH.2025.3574943

[90] Yutong Liu, Qingquan Sun, and Dhruvi Rajeshkumar Kapadia. 2025. Integrating Large Language Models into Robotic Autonomy: A Review of Motion, Voice, and Training Pipelines. *AI* 6, 7 (July 2025), 158. doi:10.3390/ai6070158

[91] Jia-Hsun Lo, Han-Pang Huang, Yen-Ching Chen, and Jen-Hau Chen. 2025. Memory Robot Design: A New Perspective From Human Brain Model and Large Language Model. *IEEE Access* 13 (2025), 28539–28549. doi:10.1109/ACCESS.2025.3538889

[92] Jia-Hsun Lo, Han-Pang Huang, and Jie-Shih Lo. 2025. LLM-based Robot Personality Simulation and Cognitive System. *Scientific Reports* 15, 1 (May 2025), 16993. doi:10.1038/s41598-025-01528-8

[93] Karthik Mahadevan, Jonathan Chien, Noah Brown, Zhuo Xu, Carolina Parada, Fei Xia, Andy Zeng, Leila Takayama, and Dorsa Sadigh. 2024. Generative Expressive Robot Behaviors using Large Language Models. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '24)*. Association for Computing Machinery, New York, NY, USA, 482–491. doi:10.1145/3610977.3634999

[94] Karthik Mahadevan, Blaine Lewis, Jiannan Li, Bilge Mutlu, Anthony Tang, and Tovi Grossman. 2025. ImageInThat: Manipulating Images to Convey User Instructions to Robots. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 757–766. doi:10.1109/HRI61500.2025.10974179

[95] Elena Malnatsky and Mike E.U. Ligthart. 2025. Fitting Humor: Age-Based Personalization for Shaping Relatable Child-Robot Interactions. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 331–341. doi:10.1109/HRI61500.2025.

10974165

[96] Vivek Mannava, Alex Mitrevski, and Paul G. Plöger. 2024. Exploring the Suitability of Conversational AI for Child-Robot Interaction. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1821–1827. doi:10.1109/RO-MAN60168.2024.10731435

[97] Antonio Martin-Ozimek, Isuru Jayarathne, Su Larb Mon, and Jouhyeong Chew. 2025. Learning Nonverbal Cues in Multiparty Social Interactions for Robotic Facilitators. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 1042–1046. doi:10.1109/HRI61500.2025.10974068

[98] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M. Jonker, and Myrthe L. Tielman. 2024. A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges. *ACM J. Responsib. Comput.* 1, 4, Article 26 (Nov. 2024), 45 pages. doi:10.1145/3696449

[99] Chinmaya Mishra, Rinus Verdonschot, Peter Hagoort, and Gabriel Skantze. 2023. Real-Time Emotion Generation in Human-Robot Dialogue Using Large Language Models. *Frontiers in Robotics and AI* 10 (Dec. 2023). doi:10.3389/frobt.2023.1271610

[100] Alireza Mortezapour and Giuliana Vitiello. 2025. Human-centered AI with focus on Human-robot interaction (Book chapter). arXiv:2507.04095 [cs.HC] https://arxiv.org/abs/2507.04095

[101] R.R. Murphy. 2001. Introduction to AI Robotics. *Industrial Robot: An International Journal* 28, 3 (June 2001), 266–267. doi:10.1108/ir.2001.28.3.266.1

[102] R.R. Murphy. 2004. Human-Robot Interaction in Rescue Robotics. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 34, 2 (May 2004), 138–153. doi:10.1109/TSMCC.2004.826267

[103] Alice Nardelli, Giacomo Maccagni, Federico Minutoli, Antonio Sgorbissa, and Carmine Recchiuto. 2025. Towards Intuitive Interaction: Cognitive Architecture for Artificial Personality, Emotional Intelligence, and Cognitive Capabilities. *International Journal of Social Robotics* 17, 10 (Oct. 2025), 2211–2228. doi:10.1007/s12369-025-01260-3

[104] Massimiliano Nigro, Emmanuel Akinrintoyo, Nicole Salomons, and Micol Spitale. 2025. Social Group Human-Robot Interaction: A Scoping Review of Computational Challenges. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 468–478. doi:10.1109/HRI61500.2025.10973980

[105] OpenAI. 2025. Introducing GPT-5. https://openai.com/index/introducing-gpt-5/. Accessed: 2025-09-05.

[106] OpenAI et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[107] Akhil Padmanabha, Jessie Yuan, Janavi Gupta, Zulekha Karachiwalla, Carmel Majidi, Henny Admoni, and Zackory Erickson. 2024. VoicePilot: Harnessing LLMs as Speech Interfaces for Physically Assistive Robots. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 116, 18 pages. doi:10.1145/3654777.3676401

[108] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. 2021. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* 372 (March 2021), n71. doi:10.1136/bmj.n71

[109] Ziqi Pan, Xiucheng Zhang, Zisu Li, Zhenhui Peng, Mingming Fan, and Xiaojuan Ma. 2025. ACKnowledge: A Computational Framework for Human Compatible Affordance-based Interaction Planning in Real-world Contexts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 892, 20 pages. doi:10.1145/3706598.3713791

[110] Giuseppe Paolo, Jonas Gonzalez-Billandon, and Balázs Kégl. 2024. A call for embodied AI. arXiv:2402.03824 [cs.AI] https://arxiv.org/abs/2402.03824

[111] Max Pascher, Uwe Gruenefeld, Stefan Schneegass, and Jens Gerken. 2023. How to Communicate Robot Motion Intent: A Scoping Review. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 409, 17 pages. doi:10.1145/3544548.3580857

[112] Andre Pereira, Lubos Marcinek, Jura Miniota, Sofia Thunberg, Erik Lagerstedt, Joakim Gustafson, Gabriel Skantze, and Bahar Irfan. 2024. Multimodal User Enjoyment Detection in Human-Robot Conversation: The Power of Large Language Models. In *Proceedings of the 26th International Conference on Multimodal Interaction* (San Jose, Costa Rica) *(ICMI '24)*. Association for Computing Machinery, New York, NY, USA, 469–478. doi:10.1145/3678957.3685729

[113] Francisco Perella-Holfeld, Samar Sallam, Julia Petrie, Randy Gomez, Pourang Irani, and Yumiko Sakamoto. 2024. Parent and Educator Concerns on the Pedagogical Use of AI-Equipped Social Robots. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 120 (Sept. 2024), 34 pages. doi:10.1145/3678556

[114] Kaitlynn Taylor Pineda, Ethan Brown, and Chien-Ming Huang. 2025. "See You Later, Alligator": Impacts of Robot Small Talk on Task, Rapport, and Interaction Dynamics in Human-Robot Collaboration. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 819–828. doi:10.1109/HRI61500.2025.10973942

[115] Maria J. Pinto and Tony Belpaeme. 2024. Predictive Turn-Taking: Leveraging Language Models to Anticipate Turn Transitions in Human-Robot Dialogue. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, 1733–1738. doi:10.1109/RO-MAN60168.2024.10731379

[116] Maria Pinto-Bernal, Matthijs Biondina, and Tony Belpaeme. 2025. Designing Social Robots with LLMs for Engaging Human Interaction. *Applied Sciences* 15, 11 (Jan. 2025), 6377. doi:10.3390/app15116377

[117] Aniket Pramanick, Yufang Hou, Saif M. Mohammad, and Iryna Gurevych. 2024. Transforming Scholarly Landscapes: Influence of Large Language Models on Academic Fields beyond Computer Science. arXiv:2409.19508 [cs.CL] https://arxiv.org/abs/2409.19508

[118] XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences* 63, 10 (Sept. 2020), 1872–1897. doi:10.1007/s11431-020-1647-3

[119] Nadun Ranasinghe, Wael M. Mohammed, Kostas Stefanidis, and Jose L. Martinez Lastra. 2025. Large Language Models in Human-Robot Collaboration With Cognitive Validation Against Context-Induced Hallucinations. *IEEE Access* 13 (2025), 77418–77430. doi:10.1109/ACCESS.2025.3565918

[120] Samira Rasouli, Moojan Ghafurian, and Kerstin Dautenhahn. 2025. Co-Design and User Evaluation of a Robotic Mental Well-Being Coach to Support University Students' Public Speaking Anxiety. *ACM Trans. Comput.-Hum. Interact.* 32, 3, Article 24 (June 2025), 70 pages. doi:10.1145/3718084

[121] Merle M. Reimann, Koen V. Hindriks, Florian A. Kunneman, Catharine Oertel, Gabriel Skantze, and Iolanda Leite. 2025. What Can You Say to a Robot? Capability Communication Leads to More Natural Conversations. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 708–716. doi:10.1109/HRI61500.2025.10974151

[122] Laurel D. Riek. 2012. Wizard of Oz studies in HRI: a systematic review and new reporting guidelines. *J. Hum.-Robot Interact.* 1, 1 (July 2012), 119–136. doi:10.5898/JHRI.1.1.Riek

[123] Diego Rodríguez-Guerra, Gorka Sorrosal, Itziar Cabanes, and Carlos Calleja. 2021. Human-Robot Interaction Review: Challenges and Solutions for Modern Industrial Environments. *IEEE Access* 9 (2021), 108557–108578. doi:10.1109/ACCESS.2021.3099287

[124] Julia Rosén, Jessica Lindblom, Maurice Lamb, and Erik Billing. 2024. Previous Experience Matters: An in-Person Investigation of Expectations in Human–Robot Interaction. *International Journal of Social Robotics* 16, 3 (March 2024), 447–460. doi:10.1007/s12369-024-01107-3

[125] Matthew Rueben, Shirley A. Elprama, Dimitrios Chrysostomou, and An Jacobs. 2020. Introduction to (Re)Using Questionnaires in Human-Robot Interaction Research. In *Human-Robot Interaction: Evaluation Methods and Their Standardization*, Céline Jost, Brigitte Le Pévédic, Tony Belpaeme, Cindy Bethel, Dimitrios Chrysostomou, Nigel Crook, Marine Grandgeorge, and Nicole Mirnig (Eds.). Springer International Publishing, Cham, 125–144. doi:10.1007/978-3-030-42307-0_5

[126] Yuki Sakamoto, Takahisa Uchida, and Hiroshi Ishiguro. 2025. Effectiveness of Conversational Robots Capable of Estimating and Modeling User Values. *International Journal of Social Robotics* 17, 6 (June 2025), 1003–1017. doi:10.1007/s12369-025-01258-x

[127] Ahmed Salem and Kaoru Sumi. 2024. A Comparative Human-Robot Interaction Study between Face-Display and an Advanced Social Robot. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 628–633. doi:10.1109/COMPSAC61105.2024.00090

[128] Sahar Salimpour, Lei Fu, Kajetan Rachwał, Pascal Bertrand, Kevin O'Sullivan, Robert Jakob, Farhad Keramat, Leonardo Militano, Giovanni Toffetti, Harry Edelman, and Jorge Peña Queralta. 2025. Towards Embodied Agentic AI: Review and Classification of LLM- and VLM-Driven Robot Autonomy and Interaction. arXiv:2508.05294 [cs.RO] https://arxiv.org/abs/2508.05294

[129] Stefan Schaal. 2007. The New Robotics—towards Human-centered Machines. *HFSP Journal* 1, 2 (July 2007), 115–126. doi:10.2976/1.2748612

[130] Benjamin Schnitzer, Umut Can Vural, Bastian Schnitzer, Muhammad Usman Sardar, Oren Fuerst, and Oliver Korn. 2024. Prototyping a Zoomorphic Interactive Robot Companion with Emotion Recognition and Affective Voice Interaction for Elderly People. *Proc. ACM Hum.-Comput. Interact.* 8, EICS, Article 242 (June 2024), 32 pages. doi:10.1145/3660244

[131] Tim Schreiter, Jens V. Rüppel, Rishi Hazra, Andrey Rudenko, Martin Magnusson, and Achim J. Lilienthal. 2025. Evaluating Efficiency and Engagement in Scripted and LLM-enhanced Human-Robot Interactions. arXiv:2501.12128 [cs] doi:10.48550/arXiv.2501.12128

[132] Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. 2025. Large VLM-based Vision-Language-Action Models for Robotic Manipulation: A Survey. arXiv:2508.13073 [cs.RO] https://arxiv.org/abs/2508.13073

[133] Jocelyn Shen, Audrey Lee, Sharifa Alghowinem, River Adkins, Cynthia Breazeal, and Hae Won Park. 2025. Social Robots as Social Proxies for Fostering Connection and Empathy towards Humanity. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI '25)*. IEEE, Melbourne, Australia, 989–999. doi:10.1109/HRI61500.2025.10973992

[134] Thomas B. Sheridan. 2016. Human–Robot Interaction: Status and Challenges. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 58, 4 (2016), 525–532. doi:10.1177/0018720816644364

[135] Zhonghao Shi, Ellen Landrum, Amy O'Connell, Mina Kian, Leticia Pinto-Alva, Kaleen Shrestha, Xiaoyuan Zhu, and Maja J Matarić. 2024. How Can Large Language Models Enable Better Socially Assistive Human-Robot Interaction: A Brief Survey. *Proceedings of the AAAI Symposium Series* 3, 1 (May 2024), 401–404. doi:10.1609/aaaiss.v3i1.31245

[136] Hirokazu Shirado, Kye Shimizu, Nicholas A Christakis, and Shunichi Kasahara. 2025. Realism Drives Interpersonal Reciprocity but Yields to AI-Assisted Egocentrism in a Coordination Experiment. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 693, 21 pages. doi:10.1145/3706598.3713371

[137] Neeraj Shrivastava, Pushpa Tewari, S. Sujatha, Srinivasa Rao Bogireddy, Neeraj Varshney, and Vinod Sharma. 2025. Natural Language Processing for Conversational AI: Chatbots and Virtual Assistants. In *2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, Vol. 3. IEEE, 1–6. doi:10.1109/IATMSI64286.2025.10984818

[138] Bruno Siciliano and Oussama Khatib (Eds.). 2008. *Springer Handbook of Robotics*. Springer, Berlin.

[139] Thomas Sievers. 2025. A Humanoid Social Robot as a Teaching Assistant in the Classroom. arXiv:2508.05646 [cs.HC] https://arxiv.org/abs/2508.05646

[140] Thomas Sievers and Nele Russwinkel. 2024. Interacting with a Sentimental Robot – Making Human Emotions tangible for a Social Robot via ChatGPT*. In *2024 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. IEEE, 182–187. doi:10.1109/ARSO60199.2024.10557749

[141] Thomas Sievers and Nele Russwinkel. 2024. Introducing a Note of Levity to Human-Robot Interaction with Dialogs Containing Irony, Sarcasm and Jocularity. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, 763–768. doi:10.1109/RO-MAN60168.2024.10731234

[142] Gabriel Skantze and Bahar Irfan. 2025. Applying General Turn-taking Models to Conversational Human-Robot Interaction. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 859–868. doi:10.1109/HRI61500.2025.10973958

[143] Micol Spitale, Minja Axelsson, and Hatice Gunes. 2025. VITA: A Multi-Modal LLM-Based System for Longitudinal, Autonomous and Adaptive Robotic Mental Well-Being Coaching. *ACM Transactions on Human-Robot Interaction* 14, 2 (March 2025), 1–28. doi:10.1145/3712265

[144] Annika Stampf, Mark Colley, Bettina Girst, and Enrico Rukzio. 2024. Exploring Passenger-Automated Vehicle Negotiation Utilizing Large Language Models for Natural Interaction. In *Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Stanford, CA, USA) *(AutomotiveUI '24)*. Association for Computing Machinery, New York, NY, USA, 350–362. doi:10.1145/3640792.3675725

[145] Carson Stark, Bohkyung Chun, Casey Charleston, Varsha Ravi, Luis Pabon, Surya Sunkari, Tarun Mohan, Peter Stone, and Justin Hart. 2024. Dobby: A Conversational Service Robot Driven by GPT-4. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, 1362–1369. doi:10.1109/RO-MAN60168.2024.10731375

[146] Ruth Stock-Homburg. 2022. Survey of Emotions in Human–Robot Interactions: Perspectives from Robotic Psychology on 20 Years of Research. *International Journal of Social Robotics* 14, 2 (March 2022), 389–411. doi:10.1007/s12369-021-00778-6

[147] Carl Strathearn and Minhua Ma. 2020. Modelling User Preference for Embodied Artificial Intelligence and Appearance in Realistic Humanoid Robots. *Informatics* 7, 3 (Sept. 2020), 28. doi:10.3390/informatics7030028

[148] Zhidong Su and Weihua Sheng. 2024. ChatAdp: ChatGPT-powered Adaptation System for Human-Robot Interaction. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5512–5518. doi:10.1109/ICRA57147.2024.10611520

[149] Fuze Sun, Lingyu Li, Shixiangyue Meng, Xiaoming Teng, Terry R. Payne, and Paul Craig. 2025. Integrating emotional intelligence, memory architecture, and gestures to achieve empathetic humanoid robot interaction in an educational setting. arXiv:2505.19803 [cs.RO] https://arxiv.org/abs/2505.19803

[150] Ryo Suzuki, Adnan Karim, Tian Xia, Hooman Hedayati, and Nicolai Marquardt. 2022. Augmented Reality and Robotics: A Survey and Taxonomy for AR-enhanced Human-Robot Interaction and Robotic Interfaces. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 553, 33 pages. doi:10.1145/3491102.3517719

[151] Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2022. CommonsenseQA 2.0: Exposing the Limits of AI through Gamification. arXiv:2201.05320 [cs.CL] https://arxiv.org/abs/2201.05320

[152] Yiliu Tang, Jason Situ, Andrea Yaoyun Cui, Mengke Wu, and Yun Huang. 2025. LLM Integration in Extended Reality: A Comprehensive Review of Current Trends, Challenges, and Future Perspectives. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1054, 24 pages. doi:10.1145/3706598.3714224

[153] Yiran Tao, Jehan Yang, Dan Ding, and Zackory Erickson. 2025. LAMS: LLM-Driven Automatic Mode Switching for Assistive Teleoperation. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 242–251. doi:10.1109/HRI61500.2025.10974127

[154] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL] https://arxiv.org/abs/2302.13971

[155] Yosuke Tsushima, Shu Yamamoto, Ankit A Ravankar, Jose Victorio Salazar Luces, and Yasuhisa Hirata. 2025. Task Planning for a Factory Robot Using Large Language Model. *IEEE Robotics and Automation Letters* 10, 3 (March 2025), 2383–2390. doi:10.1109/LRA.2025.3531153

[156] Louise Veling and Conor McGinn. 2021. Qualitative Research in HRI: A Review and Taxonomy. *International Journal of Social Robotics* 13, 7 (Nov. 2021), 1689–1709. doi:10.1007/s12369-020-00723-z

[157] Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. 2024. Theory of Mind Abilities of Large Language Models in Human-Robot Interaction: An Illusion?. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '24)*. Association for Computing Machinery, New York, NY, USA, 36–45. doi:10.1145/3610978.3640767

[158] Chenyang Wang, Daniel Tozadore, Barbara Bruno, and Pierre Dillenbourg. 2025. The Child-Robot Relational Norm Intervention to Promote Correct Handwriting Posture for Children. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 351–360. doi:10.1109/HRI61500.2025.10974144

[159] Chongyang Wang, Siqi Zheng, Lingxiao Zhong, Chun Yu, Chen Liang, Yuntao Wang, Yuan Gao, Tin Lun Lam, and Yuanchun Shi. 2024. PepperPose: Full-Body Pose Estimation with a Companion Robot. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 586, 16 pages. doi:10.1145/3613904.3642231

[160] Jiaqi Wang, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, Yincheng Yao, Xuan Liu, Bao Ge, and Shu Zhang. 2025. Large Language Models for Robotics: Opportunities, Challenges, and Perspectives. *Journal of Automation and Intelligence* 4, 1 (March 2025), 52–64. doi:10.1016/j.jai.2024.12.003

[161] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A Survey on Large Language Model Based Autonomous Agents. *Frontiers of Computer Science* 18, 6 (Dec. 2024), 186345. doi:10.1007/s11704-024-40231-1

[162] Mengyang Wang, Keye Yu, Yukai Zhang, and Mingming Fan. 2025. Challenges in Adopting Companion Robots: An Exploratory Study of Robotic Companionship Conducted with Chinese Retirees. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW045 (May 2025), 27 pages. doi:10.1145/3710943

[163] Xian Wang, Luyao Shen, and Lik-Hang Lee. 2025. A Systematic Review of XR-Enabled Remote Human-Robot Interaction Systems. *ACM Comput. Surv.* 57, 11, Article 273 (June 2025), 37 pages. doi:10.1145/3730574

[164] Yufei Wang, Wenting Zeng, Changzhen Liu, Zhuohan Ye, Jiawei Sun, Junxiang Ji, Zhihan Jiang, Xianyi Yan, Yongyi Wu, Yigao Wang, Dingqi Yang, Leye Wang, Daqing Zhang, Cheng Wang, and Longbiao Chen. 2024. CrowdBot: An Open-Environment Robot Management System for On-Campus Services. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 80 (May 2024), 27 pages. doi:10.1145/3659601

[165] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1800, 14 pages.

[166] Yize Wei, Nathan Rocher, Chitralekha Gupta, Mia Huong Nguyen, Roger Zimmermann, Wei Tsang Ooi, Christophe Jouffrais, and Suranga Nanayakkara. 2025. Human Robot Interaction for Blind and Low Vision People: A Systematic Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 276, 19 pages. doi:10.1145/3706598.3713438

[167] Thomas H. Weisswange, Hifza Javed, Manuel Dietrich, Malte F. Jung, and Nawid Jamali. 2026. Design Implications for Robots That Facilitate Groups—A Scoping

Review on Improving Group Interactions through Directed Robot Action. *J. Hum.-Robot Interact.* 15, 2, Article 44 (Jan. 2026), 108 pages. doi:10.1145/3777455

[168] Joel Wester, Bhakti Moghe, Katie Winkle, and Niels van Berkel. 2024. Facing LLMs: Robot Communication Styles in Mediating Health Information between Parents and Young Adults. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 497 (Nov. 2024), 37 pages. doi:10.1145/3687036

[169] Joel Wester, Henning Pohl, Simo Hosio, and Niels van Berkel. 2024. "This Chatbot Would Never...": Perceived Moral Agency of Mental Health Chatbots. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 133 (April 2024), 28 pages. doi:10.1145/3637410

[170] Graham Wilcock and Kristiina Jokinen. 2023. To Err Is Robotic; to Earn Trust, Divine: Comparing ChatGPT and Knowledge Graphs for HRI. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Busan, Korea, Republic of, 1396–1401. doi:10.1109/RO-MAN57019.2023.10309510

[171] Tom Williams. 2025. Improvising Interaction: Toward Applied Improvisation Driven Social Robotics Theory and Education. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 1140–1148. doi:10.1109/HRI61500.2025.10974230

[172] xAI. 2025. Grok 3 Beta — The Age of Reasoning Agents. https://x.ai/news/grok-3. Accessed: 2025-09-05.

[173] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, Qi Zhang, and Tao Gui. 2025. The Rise and Potential of Large Language Model Based Agents: A Survey. *Science China Information Sciences* 68, 2 (Feb. 2025), 121101. doi:10.1007/s11432-024-4222-0

[174] Shengyuan Xie, Eduardo Benitez Sandoval, Khaja Ahmed Shaik, and Francisco Cruz. 2025. Embodied Generative AI Art for Enhanced Human-Robot Interaction Through a Human-Centric LLM-Guided Robotic Arm Drawing System. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 1727–1730. doi:10.1109/HRI61500.2025.10974105

[175] Michael F. Xu and Bilge Mutlu. 2025. Exploring the Use of Robots for Diary Studies. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 174–182. doi:10.1109/HRI61500.2025.10974118

[176] Yuga Yano, Akinobu Mizutani, Yukiya Fukuda, Daiju Kanaoka, Tomohiro Ono, and Hakaru Tamukoh. 2024. Unified Understanding of Environment, Task, and Human for Human-Robot Interaction in Real-World Environments. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, 224–230. doi:10.1109/RO-MAN60168.2024.10731235

[177] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629 [cs.CL] https://arxiv.org/abs/2210.03629

[178] Yang Ye, Hengxu You, and Jing Du. 2023. Improved Trust in Human-Robot Collaboration With ChatGPT. *IEEE Access* 11 (2023), 55748–55754. doi:10.1109/ACCESS.2023.3282111

[179] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* 11, 12 (Nov. 2024). doi:10.1093/nsr/nwae403

[180] Hongqi Yu, Fei Tang, Lei Zhang, Randy Gomez, Eric Nichols, and Guangliang Li. 2024. Improving Perceived Emotional Intelligence of Embodied Chatbot Haru via Multi-Modal Interaction. In *2024 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 51–58. doi:10.1109/ROBIO64047.2024.10907584

[181] Fanlong Zeng, Wensheng Gan, Zezheng Huai, Lichao Sun, Hechang Chen, Yongheng Wang, Ning Liu, and Philip S. Yu. 2025. Large Language Models for Robotics: A Survey. arXiv:2311.07226 [cs.RO] https://arxiv.org/abs/2311.07226

[182] Alex Wuqi Zhang, Clark Kovacs, Liberto de Pablo, Justin Zhang, Maggie Bai, Sooyeon Jeong, and Sarah Sebo. 2025. Exploring Robot Personality Traits and Their Influence on User Affect and Experience. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 968–977. doi:10.1109/HRI61500.2025.10973991

[183] Alex Wuqi Zhang, Rafael Queiroz, and Sarah Sebo. 2025. Balancing User Control and Perceived Robot Social Agency through the Design of End-User Robot Programming Interfaces. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE, 899–908. doi:10.1109/HRI61500.2025.10974063

[184] Bowen Zhang and Harold Soh. 2023. Large Language Models as Zero-Shot Human Models for Human-Robot Interaction. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7961–7968. doi:10.1109/IROS55552.2023.10341488

[185] Ceng Zhang, Junxin Chen, Jiatong Li, Yanhong Peng, and Zebing Mao. 2023. Large Language Models for Human–Robot Interaction: A Review. *Biomimetic Intelligence and Robotics* 3, 4 (Dec. 2023), 100131. doi:10.1016/j.birob.2023.100131

[186] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence* 46, 8 (2024), 5625–5644. doi:10.1109/TPAMI.2024.3369699

[187] Renchi Zhang, Jesse van der Linden, Dimitra Dodou, Harleigh Seyffert, Yke Bauke Eisma, and Joost de Winter. 2025. Walk Along: An Experiment on Controlling the Mobile Robot "Spot" with Voice and Gestures. *J. Hum.-Robot Interact.* 14, 4, Article 64 (July 2025), 43 pages. doi:10.1145/3729540

[188] Tianyi Zhang, Colin Au Yeung, Emily Aurelia, Yuki Onishi, Neil Chulpongsatorn, Jiannan Li, and Anthony Tang. 2025. Prompting an Embodied AI Agent: How Embodiment and Multimodal Signaling Affects Prompting Behaviour. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 60, 25 pages. doi:10.1145/3706598.3713110

[189] Yuchong Zhang, Khaled Kassem, Zhengya Gong, Fan Mo, Yong Ma, Emma Kirjavainen, and Jonna Häkkilä. 2024. Human-centered AI Technologies in Human-robot Interaction for Social Settings. In *Proceedings of the 23rd International Conference on Mobile and Ubiquitous Multimedia* (Stockholm, Sweden) *(MUM '24)*. Association for Computing Machinery, New York, NY, USA, 501–505. doi:10.1145/3701571.3701610

[190] Yan Zhang, Tharaka Sachintha Ratnayake, Cherie Sew, Jarrod Knibbe, Jorge Goncalves, and Wafa Johal. 2025. Can you pass that tool?: Implications of Indirect Speech in Physical Human-Robot Collaboration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 781, 18 pages. doi:10.1145/3706598.3713780

[191] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2025. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL] https://arxiv.org/abs/2303.18223

[192] Zhaxizhuoma, Pengan Chen, Ziniu Wu, Jiawei Sun, Dong Wang, Peng Zhou, Nieqing Cao, Yan Ding, Bin Zhao, and Xuelong Li. 2025. AlignBot: Aligning VLM-Powered Customized Task Planning with User Reminders Through Fine-Tuning for Household Robots. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 12549–12556. doi:10.1109/ICRA55743.2025.11128775

[193] Qingxiao Zheng, Yiliu Tang, Yiren Liu, Weizi Liu, and Yun Huang. 2022. UX Research on Conversational Human-AI Interaction: A Literature Review of the ACM Digital Library. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–24. doi:10.1145/3491102.3501855

[194] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, and Lichao Sun. 2023. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. arXiv:2302.09419 [cs.AI] https://arxiv.org/abs/2302.09419

[195] Shaolin Zhu, Supryadi, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, and Deyi Xiong. 2024. Multilingual Large Language Models: A Systematic Survey. arXiv:2411.11072 [cs.CL] https://arxiv.org/abs/2411.11072

[196] Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Yankai Chen, Chunyu Miao, Hoang Nguyen, Yue Zhou, Weizhi Zhang, Liancheng Fang, Langzhou He, Yangning Li, Dongyuan Li, Renhe Jiang, Xue Liu, and Philip S. Yu. 2025. LLM-Based Human-Agent Collaboration and Interaction Systems: A Survey. arXiv:2505.00753 [cs.CL] https://arxiv.org/abs/2505.00753

[197] Weiqin Zu, Wenbin Song, Ruiqing Chen, Ze Guo, Fanglei Sun, Zheng Tian, Wei Pan, and Jun Wang. 2024. Language and Sketching: An LLM-driven Interactive Multimodal Multitask Robot Navigation Framework. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1019–1025. doi:10.1109/ICRA57147.2024.10611462

A Retrieval Keywords for Publication Trend

This appendix elaborates on the search keywords and Boolean retrieval strategies adopted in the ACM digital library for analyzing the publication trends of Large Language Models (LLMs), Human-Robot Interaction (HRI), and their interdisciplinary research (LLM-driven HRI) over the decade from 2015 to 2025. All retrieval operations were uniformly executed on September 9, 2025. The annual number of relevant publications retrieved for each category is as follows:

- **Large Language Models (LLMs)**: 2015 (15), 2016 (11), 2017 (12), 2018 (13), 2019 (10), 2020 (19), 2021 (33), 2022 (89), 2023 (1119), 2024 (5074), 2025 (5498)
- **Human–Robot Interaction (HRI)**: 2015 (384), 2016 (408), 2017 (549), 2018 (565), 2019 (384), 2020 (840), 2021 (681), 2022 (701), 2023 (908), 2024 (1206), 2025 (814)
- **LLM-driven HRI**: 2015 (1), 2016 (2), 2017 (0), 2018 (1), 2019 (1), 2020 (5), 2021 (7), 2022 (21), 2023 (108), 2024 (311), 2025 (268)

It should be noted that these retrieved publication counts are solely for trend reference purposes, and the relevance of the retrieved articles to the core research themes (LLMs, HRI, and LLM-driven HRI) remains to be further verified. The detailed temporal publication trends (including visualizations of annual volume changes) are presented in Figure 2.

Table 2: Search keywords and strategies used for literature retrieval in the ACM digital library (2015–2025).

Category	Search Keywords / Query
Large Language Models (LLMs)	“large language model" OR“LLM" OR“foundation model"
Human–Robot Interaction (HRI)	“human-robot interaction" OR “HRI" OR“human-robot collaboration" OR“HRC"
LLM-driven HRI	("large language model" OR LLM OR ChatGPT OR GPT-3 OR GPT-4) AND (robot OR robotics OR“social robot" OR“humanoid robot") AND ("human-robot interaction" OR HRI OR“human-robot collaboration" OR HRC)

B Related Works and Contributions

Table 3: Overview of representative review, survey, and meta-study works on LLMs and foundation models in robotics and HRI, including authors, publication year, research focus, and core contributions.

Author	Type	Year	Focus	Contribution
Zeng et al. [181]	Survey	2023	LLMs in Robotics	Highlights the benefits of LLMs for robotics.
Wang et al. [160]	Review	2025	LLMs in Robotics	Provides an overview of the integration of LLMs into robotic systems and tasks.
Firoozi et al. [39]	Survey	2023	Foundation Models in Robotics	Surveys the promising applications of foundation models in robotics.
Jeong et al. [64]	Survey	2024	Robotic Systems	Explores the potential impact and applicability of LLMs on robotics.
Kim et al. [72]	Survey	2024	LLMs in Intelligent Robotics	Offers detailed prompt engineering guidelines and categorizes LLMs in robotics.
Wang et al. [161]	Survey	2024	LLM-based Autonomous Agents	Proposes a review of LLM-based autonomous agents from a holistic perspective.
Liu et al. [90]	Review	2025	Robotic Autonomy	Surveys the integration of LLMs into autonomous robotics.
Guo et al. [90]	Survey	2024	LLM-based Multi-Agents	Provides a taxonomy of LLM-MA systems.
Li et al. [85]	Survey	2025	Multi-Robot Systems	Provides the first exploration of integrating LLMs into MRS.
Xi et al. [173]	Survey	2025	LLM-based Agents	Present a framework for LLM-based agents, comprising three components.
Lin et al. [88]	Survey	2024	Embodied AI	Proposes PALoop framework, an emotional logic engine based on LLMs.
Salimpour et al. [128]	Survey	2025	Embodied Agentic AI	Proposes a taxonomy of LLM integration and provides a comparative analysis.
Zhang et al. [185]	Review	2023	LLMs in HRI	Provides the first review of LLM applications in HRI across three categories and identifies three major challenges.
Shi et al. [135]	Survey	2024	LLMs in Socially Assistive HRI	The first surveys to focus on LLMs in SARs, informing the potential of LLMs.
Atuhurra [8]	Meta-study	2024	LLMs in HRI	Identifies thirteen key benefits of LLMs in HRI and three key risks.
Zou et al. [196]	Survey	2024	LLM-Based HAI and HAC	Proposes a taxonomy for LLM-HAS, indicating challenges and opportunities.

C Retrieval Databases, Queries and Screening Results

Table 4: Overview of databases, search queries, and screening results. The temporal scope of the search covers literature published between January 1, 2021 and August 1, 2025. For studies indexed in both ACM DL and IEEE Xplore, we followed a conservative consolidation rule whereby duplicates were attributed to ACM DL to avoid double-counting across venues.

Database	Search Query	Retrieved	Included
ACM DL	("large language model" OR LLM OR ChatGPT OR GPT-3 OR GPT-4) AND (robot OR robotics OR "social robot" OR "humanoid robot") AND ("human-robot interaction" OR HRI OR "human robot collaboration" OR HRC)	709	53
IEEE Xplore	same query as above	157	24
Nature	same query as above	10	2
Science Robotics	same query as above	2	0
International Journal of Social Robotics	same query as above	10	7
Computers in Human Behavior	simplified keyword-based search due to venue-specific search constraints: (1) (LLM OR ChatGPT OR GPT-3 OR GPT-4) AND robot AND ("human-robot interaction" OR HRI OR "human robot collaboration" OR HRC), and (2) "large language model" AND (robot OR robotics OR "social robot" OR "humanoid robot") AND ("human-robot interaction" OR HRI OR "human robot collaboration" OR HRC)	6	0

D Details of the Survey Literature Statistics

Table 5: References of Section 4.1. Contextual Perception and Understanding

Sense	Number	Papers
Multimodal Physical Perception		
Static and Semi-Static Context Injection	19	[32, 48, 54, 59, 65, 67, 68, 76, 87, 92, 107, 114, 124, 145, 155, 157, 176, 184, 197]
Modular Perception and Textual Abstraction	32	[4, 9, 14, 15, 17, 24, 27, 32, 37, 42, 47, 48, 53, 59, 63, 67, 69, 73, 76, 77, 81, 87, 94, 112–115, 121, 142, 143, 175, 188]
Integrated Visual-Language Reasoning	16	[4, 6, 14, 15, 42, 47, 48, 54, 77, 81, 91, 92, 94, 109, 157, 159]
Human-Oriented Understanding		
Emotional Grounding	18	[6, 12, 15, 17, 25, 48, 57, 68, 87, 92, 96, 103, 112, 133, 140, 143, 145, 180]
Task Intent Formulation	39	[4, 14, 15, 27, 32, 34, 36, 42, 46, 49, 57–59, 65, 67, 69, 75, 83, 91–94, 107, 109, 114, 121, 127, 141, 144, 145, 153, 155, 157, 158, 164, 175, 178, 187, 197]
Human Model Alignment	19	[12, 15, 32, 37, 42, 49, 54, 65, 67, 76, 103, 109, 126, 148, 155, 157, 168, 176, 184]

Table 6: References of Section 4.2. Generative and Agentic Interaction

Interaction	Number	Papers
Generative Social Communication		
Persona Adaptation and Conversational Fluidity	40	[5, 6, 11, 12, 16, 17, 24, 32, 48, 49, 53, 56, 57, 67–69, 75, 77, 81, 87, 92, 95, 96, 103, 112, 114, 124, 127, 141, 142, 144, 145, 158, 162, 168, 170, 175, 178, 182, 183]
Embodied Social Expressiveness	22	[4–6, 12, 16, 34, 54, 57, 75, 81, 87, 92, 93, 103, 115, 124, 142, 158, 175, 180, 183, 188]
Collaborative Task Co-Creation		
Task-Oriented Planning and Execution	28	[9, 14, 16, 17, 27, 36, 42, 46, 54, 59, 63, 65, 68, 75, 76, 81, 92, 94, 107, 109, 145, 153, 155, 164, 170, 178, 184, 197]
Creative Storytelling and Social Engagement	13	[4, 5, 9, 12, 17, 24, 35, 54, 56, 57, 69, 95, 133]
Proactive Agency		
Social Initiation	24	[12, 14, 25, 32, 47, 48, 53, 56–58, 81, 96, 114, 121, 126, 133, 140, 142, 145, 148, 155, 168, 176, 183]
Anticipatory Assistance	16	[14, 15, 63, 67, 81, 87, 92, 94, 103, 115, 136, 145, 148, 168, 176, 197]

Table 7: References of Section 4.3. Iterative Optimization and Alignment

Alignment	Number	Papers
Longitudinal Personalization and Memory		
Sustained Personalization	15	[9, 12, 14, 15, 56, 81, 91, 92, 96, 133, 143, 145, 148, 153, 162]
Episodic Memory Integration	22	[14, 17, 48, 53, 56, 67, 77, 91, 92, 96, 103, 107, 109, 114, 124, 133, 140, 141, 145, 153, 178, 184]
Multi-Level Repair		
Behavioral Repair in Task Execution	13	[9, 14, 27, 35, 36, 65, 75–77, 81, 93, 121, 182]
Emotional Repair in Social Interaction	10	[4, 11, 12, 56, 57, 73, 93, 96, 103, 175]
Repair in Ethical and Normative Alignment	10	[32, 83, 87, 93, 96, 112, 143, 144, 155, 188]

Table 8: References of Section 5. Design Components

Components	Number	Papers
Modality and Interaction Channels		
Text	53	[5, 6, 9, 15, 16, 25, 27, 32, 36, 46, 47, 53, 54, 58, 59, 63, 65, 68, 69, 73, 75, 77, 87, 91–96, 103, 107, 109, 112, 115, 124, 126, 133, 136, 140, 145, 148, 155, 157, 158, 164, 168, 170, 176, 178, 180, 183, 184, 197]
Voice	71	[4–6, 9, 11, 12, 14–17, 32, 34–36, 42, 47–49, 53, 54, 56–59, 63, 67–69, 73, 75–77, 81, 83, 87, 91–93, 95, 96, 103, 107, 112–115, 121, 124, 126, 127, 133, 140, 142–145, 148, 158, 162, 164, 168, 170, 175, 176, 178, 180, 182, 183, 187, 188, 197]
Visuals	56	[5, 6, 9, 14, 16, 17, 24, 27, 32, 34, 35, 37, 42, 46–48, 54, 56, 59, 63, 65, 68, 75, 83, 87, 92–94, 96, 103, 109, 112, 113, 127, 133, 136, 140–145, 148, 153, 155, 157, 159, 164, 170, 175, 176, 180, 182, 187, 188, 197]
Motion	52	[4, 6, 9, 11, 14, 16, 17, 25, 27, 32, 34, 37, 46, 53, 54, 57–59, 65, 69, 73, 75, 76, 81, 83, 87, 91–95, 103, 107, 109, 114, 121, 124, 126, 133, 136, 140, 141, 145, 153, 155, 157, 159, 164, 175, 178, 180, 188]
Hybrid	53	[4–6, 9, 15–17, 32, 34–37, 42, 47, 48, 53, 54, 56, 59, 63, 65, 67–69, 75–77, 81, 83, 87, 91, 92, 94, 95, 103, 112, 114, 115, 124, 127, 133, 140, 145, 148, 155, 158, 159, 162, 164, 178, 187, 188, 197]
Tangible and Haptic Interaction	9	[14, 24, 25, 35, 54, 140, 153, 162, 183]
Proximity	13	[32, 36, 49, 75, 76, 103, 114, 136, 141, 145, 159, 184, 188]
Robot Mophology		
Humanoid	39	[5, 6, 9, 11, 12, 32, 35, 37, 47–49, 53, 56, 57, 63, 67, 69, 77, 81, 91, 92, 95, 96, 103, 112, 115, 121, 124, 126, 127, 140–143, 158, 159, 168, 170, 182]
Functional	31	[14, 15, 24, 25, 27, 36, 42, 46, 58, 59, 65, 68, 73, 75, 76, 83, 87, 93, 94, 107, 114, 145, 148, 153, 155, 157, 176, 178, 183, 184, 197]
Zoomorphic	2	[164, 187]
Desktop Companions	9	[4, 17, 54, 109, 113, 133, 162, 175, 180]
VR/AR-based	5	[16, 34, 136, 144, 188]
Levels of Autonomy		
Full Autonomy	46	[5, 9, 11, 12, 15, 24, 32, 34, 35, 47, 48, 56, 65, 67, 69, 75, 77, 81, 91, 92, 95, 103, 109, 112, 114, 121, 124, 127, 133, 136, 140–143, 145, 148, 158, 159, 162, 164, 168, 170, 176, 180, 184, 197]
Semi-Autonomy	37	[4, 6, 14, 16, 17, 25, 27, 36, 37, 42, 46, 49, 53, 54, 57, 59, 63, 68, 73, 76, 83, 87, 93, 94, 96, 107, 113, 115, 126, 144, 155, 157, 175, 178, 182, 183, 188]
Teleoperation	3	[58, 153, 187]

Table 9: References of Section 6.1. Methodology

Methods	**Number**	**Papers**
Laboratory Experiment	58	[4–6, 9, 12, 14–16, 27, 32, 34–37, 42, 46–49, 59, 65, 67–69, 73, 75–77, 81, 83, 91, 94, 96, 103, 107, 112, 114, 121, 124, 127, 136, 140–145, 148, 155, 157, 170, 178, 180, 183, 184, 187, 188, 197]
Field Deployments	17	[17, 24, 25, 48, 53, 56, 58, 63, 76, 115, 133, 143, 155, 158, 164, 175, 176]
Interviews	29	[5, 11, 25, 34, 42, 54, 56–59, 63, 68, 69, 81, 83, 87, 109, 114, 124, 133, 143, 145, 153, 159, 162, 168, 175, 187, 188]
Questionnaires	70	[4–6, 9, 11, 12, 15, 17, 27, 32, 34, 35, 42, 46, 47, 49, 53, 54, 59, 65, 67–69, 73, 75, 77, 81, 83, 87, 92–96, 103, 107, 109, 112–114, 121, 124, 127, 133, 136, 140–145, 148, 153, 155, 157–159, 162, 164, 168, 175, 176, 178, 180, 182–184, 187, 188, 197]
Technical Evaluation	52	[5, 9, 14–17, 27, 32, 35–37, 42, 46–48, 53, 58, 59, 63, 65, 67–69, 75–77, 83, 92–94, 96, 103, 109, 112, 115, 121, 126, 136, 140, 143, 148, 153, 155, 159, 164, 170, 176, 178, 180, 183, 187, 197]
Wizard-of-Oz	8	[6, 56, 58, 109, 113, 126, 144, 188]
Case studies	11	[5, 32, 63, 77, 94, 103, 109, 148, 155, 157, 184]
Simulations	14	[32, 65, 68, 92–94, 103, 109, 155, 159, 164, 178, 188, 197]
Co-design workshops	6	[11, 54, 56, 57, 95, 158]
BodyStorming	2	[11, 95]
Think-aloud protocols	2	[54, 83]

Table 10: References of Section 6.2. Evaluation Strategies

Strategies	**Number**	**Papers**
Objective		
Task Efficiency and Timing	46	[4, 9, 12, 16, 25, 32, 34, 35, 42, 47–49, 56, 59, 65, 68, 69, 73, 76, 77, 81, 83, 87, 92, 94, 103, 107, 115, 121, 124, 126, 133, 136, 140, 142, 153, 155, 159, 164, 175, 176, 178, 180, 187, 188, 197]
Task Accuracy and Performance	42	[14–16, 27, 32, 36, 37, 42, 47, 54, 59, 63, 65, 68, 73, 75, 77, 83, 87, 91, 92, 94, 96, 103, 109, 112, 114, 126, 133, 136, 140, 145, 148, 155, 157–159, 164, 176, 178, 184, 187]
LLM-Specific Performance	30	[15–17, 32, 36, 46–48, 53, 54, 59, 65, 68, 69, 91–93, 95, 103, 112, 113, 115, 126, 155, 157, 168, 170, 178, 180, 184]
Subjective		
User's Perceptual and Relational Experience	65	[4–6, 11, 12, 15, 17, 25, 32, 34–36, 47, 49, 53, 54, 56–59, 63, 67, 69, 75, 77, 81, 83, 87, 93–96, 103, 109, 112–114, 121, 124, 126, 127, 133, 136, 140, 141, 143–145, 148, 153, 155, 157–159, 162, 168, 175, 176, 178, 180, 182–184, 188, 197]
Perceived Intelligence	30	[14, 15, 24, 32, 34, 35, 42, 47, 54, 56, 58, 63, 67, 69, 81, 83, 92, 103, 113, 124, 126, 127, 133, 140, 143, 145, 168, 180, 183, 197]
Anthropomorphism	19	[4, 5, 9, 12, 24, 32, 34, 42, 53, 67, 69, 92, 126, 127, 133, 140, 143, 145, 180]
Usability	31	[14, 24, 27, 32, 42, 54, 58, 59, 68, 73, 76, 81, 83, 87, 94, 103, 107, 109, 113, 121, 141, 143–145, 153, 155, 158, 164, 175, 183, 187]
Safety	24	[14, 15, 34, 36, 42, 47, 54, 57, 67, 69, 73, 76, 91, 96, 113, 127, 133, 136, 143, 144, 148, 155, 162, 184]
Cognitive Load and Workload	13	[32, 42, 49, 54, 81, 83, 94, 124, 143, 153, 158, 178, 187]

Table 11: References of Section 7. Applications

Applications	**Number**	**Papers**
Social and Conversational Systems	18	[5, 32, 37, 48, 92, 93, 103, 112, 115, 126, 127, 133, 140–142, 157, 180, 184]
Healthcare and Wellbeing	12	[11, 17, 56, 57, 76, 87, 91, 107, 143, 148, 168, 182]
Domestic and Everyday Use	17	[6, 14, 24, 25, 46, 59, 65, 73, 81, 94, 109, 124, 159, 162, 175, 183, 197]
Teaching and Education	13	[4, 9, 35, 49, 54, 63, 67, 77, 83, 95, 96, 113, 158]
Industrial Manufacturing	7	[15, 36, 68, 75, 114, 155, 178]
AR/VR-enabled Interactions	6	[16, 34, 136, 144, 187, 188]
Public Space Service	9	[42, 47, 53, 58, 121, 145, 164, 170, 176]
Other	4	[12, 27, 69, 153]