

Why did My Robot Just Change Personality?

Prompting Guidelines for a Grounded Robot Persona in LLM-Based HRI

Ashita Ashok¹, Franziska Babel², Patrick Holthaus³, Rucha Khot⁴, Karla Bransky⁵, Fethiye Irmak Dogan⁶,
Karsten Berns¹, Silvia Rossi⁷, Minha Lee⁴, Guy Laban^{8,9,10*}

Abstract—Large language models (LLMs) are increasingly used for verbal interaction in social robots, yet prompt design in human-robot interaction (HRI) remains underspecified. As a result, robots may present hallucinated capabilities, unclear behavioural boundaries, and misleading personas. This paper develops a framework for prompt design in LLM-based robots and introduces a structured prompt template comprising eight functional components through which robot behaviour can be specified, bounded, and adapted. The framework is grounded in a review of prior LLM-based HRI work and complemented by survey and discussion data from HRI experts gathered at the Robo-Identity workshop at IEEE RO-MAN 2025 ($N = 27$). The qualitative findings highlight limited legibility of robot personality, the need for user adaptation, and strong ethical concerns about safety, deception, and governance. Based on these findings, we present prompting guidelines accompanied by proof-of-concept template as a structured design and reporting aid for HRI research. We argue that prompt design should be treated as a socio-technical problem rather than a minor implementation detail, requiring explicit capability boundaries, transparent behavioural assumptions, and context-sensitive safeguards to support reliable and interpretable HRI.

I. INTRODUCTION

Social interactive robots, particularly those with a human-like appearance, increasingly rely on verbal interaction powered by large language models (LLMs) [1], [2]. These models can support natural language conversation and provide access to broad world knowledge [1], but they are also prone to hallucination [3] and remain fundamentally *disembodied* from the robot’s hardware, environment, and user expression. In human-robot interaction (HRI), this gap can produce fluent but misleading interaction, for example, when a robot implies knowledge, perception, or capabilities that it does not actually possess [4]. Grounding robot behaviour in its actual capabilities and interaction context is therefore critical for calibrating user expectations and trust [5]. As LLMs increasingly shape robot communication, prompt design becomes

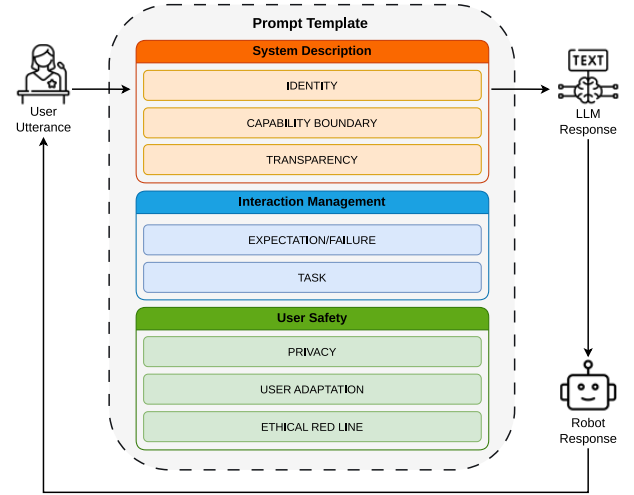


Fig. 1. A conceptual framework for prompt design in LLM-powered robots.

a central HRI problem rather than a minor implementation detail: before interaction begins, LLM-based robots must be configured through structured prompt components that define the robot’s identity, capabilities, disclosures, and behaviour under uncertainty.

Accordingly, this paper aims to clarify *what* should be specified in such prompts and *why*. We pursue this goal in two complementary steps. First, we draw on broader HRI scholarship to develop a theoretical framework for understanding how robot identity, behaviour, expectations, transparency, and behavioural boundaries are currently constructed and interpreted in embodied interaction, and how these may shape prompt design as a central mechanism of robot behaviour and identity. We then use this framework as an analytic lens to survey selected recent LLM-based HRI studies and identify the core prompt components through which these aspects are specified, bounded, or left implicit in practice, leading to a recommended prompt template for HRI. Second, we complement this framework and literature-derived template with an empirical layer based on an expert survey and discussion from three HRI expert groups during the workshop “Robo-Identity : Methods & Theories in Personification of Social Robots” at RO-MAN 2025 with HRI researchers and practitioners. Together, these two steps allow us to address three linked challenges in LLM-based HRI: supporting a consistent and interpretable robot identity and behaviour s [6], [7], reducing unnecessary deception and misleading interaction [8], [9], and limiting

*Corresponding author: Guy Laban - Laban@bgu.ac.il

¹Robotics Research Lab, RPTU Kaiserslautern, Kaiserslautern, Germany. Email: ashita.ashok, karsten.berns@cs.rptu.de; ²Human-Centered Systems, Linköping University, Linköping, Sweden. Email: franziska.babel@liu.se; ³Robotics Research Group, University of Hertfordshire, Hatfield, UK. Email: p.holthaus@herts.ac.uk; ⁴Industrial Design, Eindhoven University of Technology, Eindhoven, Netherlands. Email: {r.khot,m.lee}@tue.nl; ⁵Gameflow Lab, Australian National University, Canberra, Australia. Email: karla.kelly@anu.edu.au; ⁶Affective Intelligence and Robotics Laboratory, University of Cambridge, Cambridge, UK. Email: fid21@cam.ac.uk; ⁷Intelligent Robotics and Advanced Cognitive System Projects Laboratory, University of Naples Federico II, Naples, Italy. Email: silvia.rossi@unina.it; ⁸LabaLab - Language, Affect, and Behaviour in AI Lab, Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer Sheva, Israel. Email: laban@bgu.ac.il; ⁹School of Brain Sciences and Cognition, Ben-Gurion University of the Negev, Beer Sheva, Israel; ¹⁰The Azrieli National Center for Autism and Neurodevelopment Research, Beer Sheva, Israel.

hallucinations and capability mismatches in embodied settings [4], [10]. The outcome of this work is a structured prompt template comprising eight functional components (see Fig 1): IDENTITY, CAPABILITY BOUNDARY, TRANSPARENCY, TASK, EXPECTATION/FAILURE PROTOCOL, PRIVACY, USER-ADAPTATION, and ETHICAL RED LINES. We present this template as a structured design and reporting aid for HRI research rather than a fixed recipe for all robots and contexts. In this way, the paper contributes a framework for defining bounded artificial robot personas that align with system capabilities, make prompt assumptions explicit, and support more transparent, comparable, and reproducible LLM-based HRI.

II. THEORETICAL FRAMEWORK

Prompt design can be understood as a central mechanism through which robot behaviour is configured for interaction. Rather than treating behaviour as an emergent product of language generation alone, this perspective highlights that prompts shape how a robot is presented, what it is expected to do, and how its conduct is constrained in social interaction. Drawing on prior HRI scholarship on robot identity, expectation formation, transparency, and interaction failure, we use this framework to examine how prompt design structures artificial personas and their boundaries in embodied HRI.

a) Identity: Identity in HRI is not peripheral but central, emerging through design strategies, social dynamics, and users' own constructions of robot personas. Many aspects contribute to perceived identity, including embodiment and behavioural cues [7]. Users rely on stereotype projection based on appearance (e.g., gender, ethnicity, or racial bias) [11], while hybrid embodiments further shape perception [7], for example, when additional interfaces increase perceived competence [12]. Research in Robo-Identity has further highlighted that identity should be understood as a socially constructed and evolving phenomenon [13]–[16], shaped over time through interaction and context [11]. Prior work on artificial identity design also outlines key principles for specifying identity attributes, including role, persona, embodiment, and functional constraints, but does not operationalise these for LLM-based systems [7].

Operationalising identity in prompts therefore requires explicit specification of a coherent set of characteristics. This includes the robot's role and persona (e.g., communication style, expressiveness, and behavioural tone) [7], as well as constraints derived from embodiment and system capabilities. Prior work highlights that identity is closely tied to embodiment, system design, and user perception [7], [17], requiring designers to define how an agent is presented and understood in context. Therefore, robot identity should be treated as a designed yet evolving construct that maintains behavioural consistency while remaining aligned with embodiment, capabilities, and interaction context.

b) Transparency: Transparency by Design in HRI points to behavioural strategies that make a robot's intentions, decisions, and internal states intelligible to human users, thereby supporting safety, predictability, trust, and

effective collaboration. Transparency is a functional requirement in contexts where humans must rapidly interpret robot behaviour and coordinate with autonomous systems. Accordingly, transparency involves making robot behaviour and internal states interpretable and legible at a communicative level [18] using signals like social gaze [19]. It includes not only how robots signal communicative intent, but also how they convey confidence, uncertainty, and progress during interaction, for example, via inner speech and emotional expression [20], [21]. Such cues help users form a more accurate understanding of what the robot is doing and why. Notably, the intention to be transparent is not the same as being transparent in an interaction. For instance, roboticists may aim to transparently convey information across different affordances like a robot verbally or visually (e.g., using a screen) describing what it is doing, but human interactants may still face high informational load, and thus interactional transparency is not delivered as intended [12].

The use of social and emotional cues introduces the ethically complex issue of deception in HRI. Although some robot behaviours and designs may, in some cases, increase compliance, likeability, or perceived agency [22]–[25], they also risk undermining trust, particularly when users interpret the deception as deliberate [26]. Systematic analysis further highlights the broader implications of deception in HRI and the need for safeguards [9]. One mitigation strategy is to incorporate behavioural cues that implicitly signal uncertainty or deception, thereby preserving a degree of transparency [27]. Overall, transparency in HRI should be understood as a multidimensional construct encompassing legibility, predictability, and the explicit communication of system limits, complicated by if and how multimodal communication, e.g., speech and GUI presentation, actually helps or hinders transparency in interactions.

c) Explanations and Expectations: Shaping robot behaviour in HRI requires acknowledging that people perceive robots holistically [28]; accordingly, explanations and expectations should be treated as a coupled design problem rather than as separate concerns. Expectations define what users believe a robot can or should do, while explanations help establish, calibrate, and revise those beliefs over time. From this perspective, explanations are not merely post-hoc accounts of robot decisions; they are part of the mechanism through which robots communicate capability, limitation, and social role [29], [30]. This is especially important in socially situated interaction, where behaviour must remain intelligible, socially appropriate, and aligned with human preferences [31].

In this context, explanations are closely tied to expectation formation and adjustment. In socially situated environments, common-sense reasoning alone is often insufficient, particularly when appropriateness depends on context or when users hold differing expectations. Prior work shows that incorporating human explanations into action generation can improve alignment with social norms and produce more interpretable robot behaviour [31]. At the same time, expectations shape how explanations are perceived, particularly in

situations of failure and recovery. Evidence from priming and failure-recovery studies suggests that expectations can significantly influence perceived robot competence, and that explanations are more effective when they support calibrated expectations [29]. These findings highlight the need to design explanations that not only account for behaviour, but also embodiment and environment to meet people’s expectations during an interaction.

III. PROMPT DESIGN IN LLM-BASED HRI

Prior work on LLM-based HRI has introduced a growing range of robot applications, but has been far less explicit about how robot behaviour is actually specified in prompts. To situate our framework, we surveyed selected recent LLM-based HRI studies from the past three years that explicitly reported prompt design (see Table I), with attention to which aspects of robot behaviour were specified, constrained, or left implicit. Given the still limited and rapidly evolving nature of this literature, we used a selected sample of representative studies to map emerging prompting practices rather than to provide an exhaustive systematic review. Our analysis was guided by one central question: which aspects of robot behaviour are being specified, bounded, or left implicit at the prompt level? This question was informed by the theoretical framework (see Section II), which treats robot persona and behaviour as the mechanism through which a robot’s identity, behavioural boundaries, transparency, interactional purpose, and user-facing safeguards are defined. Accordingly, for each paper, we examined how the prompt specified the robot’s role and persona, interaction goal, handling of uncertainty or failure, and whether it made explicit any capability limits, disclosures, privacy assumptions, user-adaptation strategies, or ethical constraints.

Most reviewed works define the robot identity within the prompt (90%) and specify the interaction task (60%). Other prompt elements appear less consistently. Expectation or failure-handling strategies are present in about half of the prompts (50%), while explicit user adaptation is rare (30%). Privacy-related instructions appear only sporadically (20%), despite several systems storing conversation histories or maintaining persistent user profiles. Explicit ethical constraints are reported in a limited subset of prompts (30%). Notably, none of the reviewed prompts specifies the robot’s embodiment or system capability boundaries (e.g., sensing limitations, mobility constraints, or knowledge access), nor do they disclose the use of LLM to the user. Consequently, both the capability boundary and transparency components remain largely unaddressed. Overall, current prompt designs in LLM-powered HRI primarily establish robot identity and task, while other aspects important for predictable and responsible interaction remain inconsistently specified.

Based on the reviewed studies, we identified eight prompt components that HRI researchers should report when documenting LLM-based robotic systems. Explicitly describing these components helps make clear how robot behaviour is defined, bounded, and adapted, while supporting transparency, comparability, and reproducibility in the field.

TABLE I
COMPARISON OF RELEVANT WORKS ON LLM-POWERED SHRS

Reference (Sample)	Model	Social Robot	Use Case
Billing et al. [32] (N=N/A)	OpenAI GPT-3	Pepper & Nao	Robot dialogue generation
Kim et al. [33] (N=32)	OpenAI GPT-3.5	Pepper	LLM capability evaluation
Verhelst et al. [34] (N=21)	OpenAI GPT-3.5	Furhat	Language tutoring
Addlesee et al. [35] (N=N/A)	Vicuna-13b-v1.5	ARI	Multi-party HRI
Kim et al. [36] (N=24)	OpenAI GPT-4o-mini (fine-tuned)	piBo	Child-centric intention-aware HRI
Skantze et al. [4] (N=39)	OpenAI TurnGPT (fine-tuned)	Furhat	Turn-taking modelling
Pinto et al. [37] (N=50)	OpenAI GPT-3.5	Pepper	Memory-aware HRI with elderly
Mauliana et al. [38] (N=13)	GoogleAI Flan-T5-Large (fine-tuned)	Ameca	Memory-aware multi-session HRI with students
Sievers et al. [39] (N=N/A)	OpenAI GPT-4o	Pepper	Knowledge retrieval assistance
Laban et al. [40] (N=21)	OpenAI GPT-3.5	QTrobot	Emotion regulation intervention

- **IDENTITY:** Defines the robot’s persona and social role assumed by the language model.
- **CAPABILITY BOUNDARY:** Specifies the robot’s perceptual inputs, embodiment constraints, and system limitations (e.g., sensing, memory, mobility, internet access).
- **TRANSPARENCY:** Specifies whether and how the robot system discloses its artificial nature or operational limitations to the user.
- **TASK:** Defines the primary conversational objective the robot should accomplish.
- **EXPECTATION/FAILURE PROTOCOL:** Specifies how the robot handles uncertainty, missing information, or interaction breakdowns.
- **PRIVACY:** Specifies constraints on the collection, storage, and use of user data by robot system.
- **USER-ADAPTATION:** Defines how the robot adapts communication to the target user group (e.g., age, language, cognitive ability).
- **ETHICAL RED LINES:** Specifies constraints preventing harmful, biased, unsafe, or illegal robot responses.

To illustrate how the identified prompt components can be operationalised in practice, we provide two **complete prompt examples** using all eight modular design components as an OSF project¹ (see Section *Sample Prompt*). An **interactive demo** is available on Hugging Face², and the corresponding **implementation** is released on GitHub³.

¹<https://osf.io/qrnzp/files/9xbpr>

²<https://huggingface.co/spaces/aashok/LLM-PSR>

³https://github.com/aashok94/llm-powered_social_robot

IV. QUALITATIVE ASSESSMENT

Building on the theoretical framework, we introduce a complementary empirical layer based on a participatory expert survey and discussion conducted at the Robo-Identity workshop. The purpose of this data collection is not to derive the framework inductively, but to empirically examine how HRI researchers and practitioners interpret its core concerns, including persona legibility, continuity over time, safety boundaries, and context-sensitive governance. As a complementary empirical layer, this study draws on data collected during a workshop at RO-MAN 2025⁴, including a survey and three focus groups. The workshop examined how artificial personas in social robots are constructed, interpreted, and discussed within the context of recent advances in LLMs and generative AI. This participatory study was approved by the Ethics Committee of the Department of Social Sciences, RPTU Kaiserslautern (Approval No. SoWi/69). Following the presentations in the workshop, participants formed three discussion groups. Two workshop organizers joined the groups to facilitate the discussions, while each group collaboratively completed a structured survey summarizing their perspectives and discussion outcomes.

All participants were experts in HRI (most > 5-9 years of experience), of different cultures and locations (Asia, Europe, US, Oceania), as well as backgrounds extending HRI (computer science, psychology, ethics, AI, social sciences). Participants were divided into three groups (Group 1: $n = 10$, Group 2: $n = 8$, Group 3: $n = 9$), resulting in a total sample of $N = 27$ participants. Each group collaboratively completed a single 12-item survey with open-ended statements designed to elicit group discussion on different aspects of robotic persona and to capture the main outcomes of the group deliberation. In addition, the discussions included two concluding free-response questions (persona adaptation for different user groups; ethical red line in persona design). Each statement included a Likert scale (1 = strongly disagree, 5 = strongly agree) and the items were designed to capture participants' perspectives on the design and interpretation of robot personas in AI-enabled social robots. The full questionnaire, including item sources and supporting literature, is available on OSF⁵. Thematic analysis was conducted at a semantic level on the data collected by two authors independently, focusing on how participants reasoned about robot persona and identity, their recognizability, continuity over time, and ethical implications.

V. RESULTS OF THEMATIC ANALYSIS

Results are presented as a qualitative-led analysis, with Likert responses used descriptively to indicate convergence and divergence across groups. Groups were mostly in agreement about most statements, with the exceptions of items 2, 3, 4, 5 and 11 (see Fig. 2).

TABLE II
THEMES FROM THE QUALITATIVE ANALYSIS.

Theme	Qualitative interpretation	Illustrative quote
Persona as a Useful but Contested Design Abstraction	Participants saw persona as technically possible, but questioned whether it is always the right or sufficiently well-defined design direction.	<i>"Yes — it can, the question, should it... we should discuss whether it is the right approach."</i> (Group 2)
Robot Personality Is Not Necessarily Legible	Participants doubted that users reliably infer stable personality from robot behavior, especially when cues are subtle and user interpretation varies.	<i>"Subtle cues are often missed."</i> (Group 2) <i>"subtleties are not clear, extremes are clear."</i> (Group 3)
Identity Over Time Was Framed as a Memory Problem	Longitudinal persona was framed as requiring memory, temporal ordering, and coherent state across interactions, rather than prompt style alone.	<i>"It can support, but would be limited in memory long term due to limited context window."</i> (Group 2)
Safety and Ethics Structured the Discussion	Participants evaluated persona and identity primarily through risks of hallucination, adversarial prompting, and deceptive or unsafe outputs.	<i>"Depends on context, deception is part of robots design."</i> (Group 2) <i>"people have died from llm suggestions."</i> (Group 3)
Context-Sensitive and Vulnerability-Aware Design	Participants rejected one-size-fits-all personas and emphasised tailoring, transparency, privacy boundaries, and greater caution in vulnerable contexts.	<i>"not a one thing fits all."</i> (Group 3) <i>"Not being transparent... using the persona of a real person..."</i> (Group 2)

A. Persona as a Useful but Contested Design Abstraction

Consistent across the responses was that participants did not reject the idea of robot persona outright, neither did they treat it as desirable. Instead, persona was positioned as a potentially useful design abstraction whose legitimacy depends on how it is defined and what work it is expected to do. Group 2 in response to item 1 acknowledged the technical feasibility of generating companion-like behaviour, but reframed the issue as normative: *"Yes — it can, the question, should it [...] we should discuss whether it is the right approach."* Group 3 expressed a similar hesitation questioning how such an approach could account for contextual variation such as culture. This tension was also visible in the ratings of items 1-3 (see Fig 2).

B. Robot Personality Is Not Necessarily Legible

A second pattern concerned the difficulty of making robot personality reliably legible. Participants suggested that subtle behavioural cues are unlikely to translate into stable interpretations across users. Group 2 stated that *"Subtle cues are often missed,"* adding that *"Neurodiversity might affect perception and recognition."* Group 3 further clarified that personality may only be recognisable under exaggerated conditions (e.g., neurotic Roomba) *"subtleties are not clear, extremes are clear,"* and that interpretation depends on user experience. This was also reflected in the ratings. Groups expressed low agreement that users can accurately recognise a robot's personality, and disagreed that LLMs can generate emotions in real time.

⁴<https://rllab.cs.rptu.de/en/conferences/ro-man-2025>

⁵<https://osf.io/qmznp/files/9xbpr>

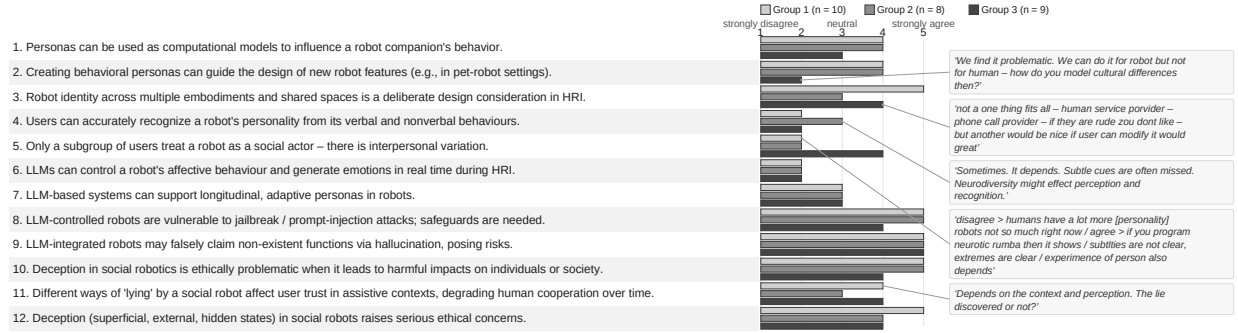


Fig. 2. Bar chart showing the agreement or disagreement to the 12 statements about LLMs and Personas presented to the 27 workshop participants in three groups. Topics: 1) Personas as Design Constructs (items 1-3); 2) Anthropomorphism and Deception (items 4-5); 3) AI-Driven Personas (6-9); 4) Deception and Ethics (items 10-13).

C. Identity Over Time Was Framed as a Memory Problem

When discussing continuity, participants framed identity as dependent on memory and system architecture rather than prompt design alone. Group 2 noted that longitudinal persona is limited by context, while Group 3 emphasised that LLMs alone cannot sustain continuity without memory of prior interactions and temporal sequencing. Participants did not reject adaptive identity, but conditioned it on the presence of mechanisms for retaining and organising interaction history.

D. Safety and Ethics Structured the Discussion

Participants consistently framed persona and identity through risks related to hallucination, manipulation, and misleading representations. Group 3 described safeguards as necessary (*"internal safeguard"*) and referenced real-world harms (*"people have died from llm suggestions"*). They also highlighted that risk extends beyond physical actions to verbal outputs. Group 2 emphasised that the impact of deception on trust depends on context and whether it is detected. The answers to statement 12, whether deception in social robots raised ethical concerns were interesting as Group 2 stated that it *"Depends on context, deception is part of robots design."* while Group 3 thought that *"We should be more careful when designing prompts with vulnerable groups; is it lying or not safeguarded; how we check the rights and wrongs."*

E. Context-Sensitive and Vulnerability-Aware Design

Participants argued against one-size-fits-all persona design and emphasised the importance of context, culture, and user vulnerability. Consistent with this, Group 3 also noted that identity across embodiments (i.e., migrating agents) is unlikely to follow a single universal model, suggesting the need for adaptation based on context or user preference. They also emphasised increased caution when designing for vulnerable groups. Participants articulated explicit boundaries. Group 2 identified lack of transparency, use of real-person personas, and harmful behaviour as unacceptable. Group 3 raised privacy concerns and emphasised the need for ethical oversight. Responses further suggested that persona may evolve over time with the user, reinforcing the need for adaptable and governed interaction design.

F. Overall Interpretation

Across the dataset, persona was seen as potentially useful, but contingent on legibility, system support for continuity including strong safety and ethical constraints. Agreement was strongest on risks related to hallucination, manipulation, and deception, while claims about stable personality perception and real-time emotional generation were met with scepticism. Disagreement centred on how broadly persona can be generalised as a design tool.

VI. DISCUSSION AND CONCLUSION

This paper contributes a structured prompt template for LLM-based robots by combining a review of prior work with a complementary empirical layer from HRI experts. The literature showed that existing systems most often specify IDENTITY and TASK, while components such as CAPABILITY BOUNDARY and TRANSPARENCY are far less consistently reported. The qualitative findings suggest that this imbalance matters. HRI experts agreed that if those components are not well defined, robot personas might not be useful as they might not be legible and adapted to the user or even deceptive and harmful to vulnerable groups.

While the literature-derived template highlights several functional prompt components, the qualitative results show that safety, deception, hallucination, and governance are among the most consequential concerns in practice, which researchers care about. Taken together, these findings suggest that the components least often reported in prior work may be among the most important for responsible deployment.

The findings also refine how robot identity could be understood. Prior HRI research shows that identity is shaped through embodiment, role, and interaction context rather than language alone [7], [12], [14], [17], [40]. Consistent with this, researchers in our sample questioned whether users can reliably perceive a stable robot personality from behaviour, especially when cues are subtle or interpreted differently across users. The qualitative layer, therefore, suggests that prompt-based identity specification is necessary, but not sufficient. Persona must also be supported by coherent behavioural cues and embodiment-consistent signalling if it is to become legible in interaction. A related contri-

bution concerns continuity over time. Participants framed longitudinal identity primarily as a memory problem rather than a prompt-style problem. This significantly extends the template. If a robot is meant to sustain a stable persona across repeated encounters, prompt design must be coordinated with memory architecture and with governance over stored user information. In practice, this means that identity design in LLM-based robots cannot be separated from questions of memory, privacy, and temporal coherence.

More broadly, the findings suggest how the community could use these guidelines. We do not present the template as a fixed recipe for all robots or contexts. Rather, it is envisioned to be used as a structured design and reporting aid that helps researchers make prompt assumptions explicit, comparable, and open to review. This is especially important in capability-limited or non-humanoid robots, where fluent language may otherwise imply abilities the system does not possess. Explicitly stating capability limits, such as a lack of physical manipulation or internet access, can help calibrate expectations and reduce misleading interaction. Future work should explore the proposed prompt guidelines implemented in different robot embodiments to evaluate the perception of the robot identity.

Overall, this paper argues that prompt design in LLM-based robots should not be treated as a minor implementation detail. The literature identifies a preliminary structure for prompt design, while the qualitative findings clarify where this structure needs stronger boundaries, clearer justification, and greater sensitivity to context. Bringing these layers together supports a more disciplined HRI approach to robot persona design, one that treats identity as inseparable from capability, transparency, continuity, and safety.

REFERENCES

- [1] G. Laban and E. S. Cross, "Sharing our Emotions with Robots: Why do we do it and how does it make us feel?" *IEEE Transactions on Affective Computing*, 2024.
- [2] Y. Kim, D. Kim, J. Choi, J. Park, N. Oh, and D. Park, "A survey on integration of large language models with intelligent robots," *Int. Service Robotics*, 2024.
- [3] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of Hallucination in Natural Language Generation," *ACM Comput. Surv.*, vol. 55, no. 12, Mar. 2023.
- [4] G. Skantze and B. Irfan, "Applying General Turn-taking Models to Conversational Human-Robot Interaction," in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2025, pp. 859–868.
- [5] J. Kraus, D. Scholz, D. Stiegemeier, and M. Baumann, "The More You Know: Trust Dynamics and Calibration in Highly Automated Driving and the Effects of Take-Overs, System Malfunction, and System Transparency," *Human factors*, vol. 62, no. 5, pp. 718–736, 2020.
- [6] W. Laity, K. Haring, and P. Holthaus, "Robot Continuity across Embodiments: Portability, Identity and Migration of Robotic Systems," in *International Conference on Robot and Human Interactive Communication*. IEEE, 2025.
- [7] K. K. Bransky, P. Sweetser, P. Holthaus, G. Laban, K. Winkle, N. Akalin, A. Ashok, J. Lee, R. Khot, A. Bejarano, J. Thijn, R. K. Moor, M. Jang, and M. Lee, "Designing Artificial Identity: The Identity Design Framework and Research Agenda," in *DIS '26: Proceedings of the 2026 ACM Designing Interactive Systems Conference*, 2026.
- [8] F. Babel, S. Saunderson, and S. Rismani, "Towards a framework for levels of anthropomorphic deception in robots and ai," in *CHI '26 Workshop: Bridge Over Troubled Water - Aligning Commercial Incentives with Ethical Design Practice to Combat Deceptive Patterns*, 2026, in press.
- [9] R. Esposito, A. Rossi, and S. Rossi, "Deception in HRI and its Implications: a Systematic Review," *ACM Transactions on Human-Robot Interaction*, 2025.
- [10] J. Atuhurra, "Leveraging Large Language Models in Human-Robot Interaction: A Critical Analysis of Potential and Pitfalls," *arXiv*, 2024.
- [11] K. Seaborn, "Social identity in human-agent interaction: A primer," *ACM Transactions on Human-Robot Interaction*, 2025.
- [12] M. Lee, P. Ruijten, L. Frank, and W. IJsselstein, "Here's looking at you, robot: the transparency conundrum in hri," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023.
- [13] M. Lee, D. Kontogiorgos, I. Torre, M. Luria, R. Tejwani, M. J. Dennis, and A. Pereira, "Robo-identity: Exploring artificial identity and multi-embodiment," in *Companion of the 2021 ACM/IEEE International Conference on HRI*, 2021.
- [14] G. Laban, S. Le Maguer, M. Lee, D. Kontogiorgos, S. Reig, I. Torre, R. Tejwani, M. J. Dennis, and A. Pereira, "Robo-identity: exploring artificial identity and emotion via speech interactions," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 1265–1268.
- [15] K. Seaborn, "From identified to self-identifying: social identity theory for socially embodied artificial agents," in *Proceedings of the HRI 2022 Workshop on Robo-Identity*, vol. 2, 2022, pp. 1–3.
- [16] R. Khot, M. Lee, A. Bejarano, L. Miranda, G. Reyes-Cruz, J. E. Fischer, and D. Kontogiorgos, "Robo-identity: Designing for identity in the shared world," in *Companion of the 2024 ACM/IEEE International Conference on HRI*, 2024.
- [17] K. Bransky, P. Sweetser, S. Caldwell, and K. Fletcher, "Mind-body-identity: A scoping review of multi-embodiment," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 65–75.
- [18] P. Holthaus, T. Schulz, G. Lakatos, and R. Soma, "Communicative Robot Signals: Presenting a New Typology for Human-Robot Interaction," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023.
- [19] F. Vigni and S. Rossi, "Exploring non-verbal strategies for initiating an hri," in *International Conference on Social Robotics*. Springer, 2022, pp. 280–289.
- [20] G. Angelopoulos, A. Rossi, C. Di Napoli, and S. Rossi, "You are in my way: non-verbal social cues for legible robot navigation behaviors," in *IROS*, 2022.
- [21] G. Angelopoulos, C. Di Martino, A. Rossi, and S. Rossi, "Unveiling the learning curve: enhancing transparency in robot's learning with inner speech and emotions," in *2023 RO-MAN*, 2023.
- [22] G. Laban, J. N. George, V. Morrison, and E. S. Cross, "Tell me more! Assessing interactions with social robots from speech," *Paladyn, Journal of Behavioral Robotics*, vol. 12, no. 1, pp. 136–159, 2021.
- [23] G. Laban, A. Kappas, V. Morrison, and E. S. Cross, "Building Long-Term Human-Robot Relationships: Examining Disclosure, Perception and Well-Being Across Time," *International Journal of Social Robotics*, vol. 16, no. 5, 2024.
- [24] G. Laban, V. Morrison, A. Kappas, and E. S. Cross, "Coping with Emotional Distress via Self-Disclosure to Robots: An Intervention with Caregivers," *International Journal of Social Robotics*, vol. 17, no. 9, pp. 1837–1870, 2025.
- [25] G. Laban, S. Chiang, and H. Gunes, "What People Share With a Robot When Feeling Lonely and Stressed and How It Helps Over Time," in *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2025, pp. 1930–1935.
- [26] A. Rossi, G. Falcone, and S. Rossi, "It is the way you lie: Effects of social robot deceptions on trust in an assistive robot," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, 2024.
- [27] R. Esposito, A. Rossi, M. Ponticorvo, and S. Rossi, "Roboleaks: Non-strategic cues for leaking deception in social robots," in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2025.
- [28] I. Torre, F. I. Dogan, and D. Kontogiorgos, "Voice, embodiment, and autonomy as identity affordances," in *HRI 2021-Robo-Identity: Exploring Artificial Identity and Multi-Embodiment March 2021*, 2021.
- [29] E. Yaddalahi, F. I. Dogan, Y. Zhang, B. Nogueira, T. Guerreiro, S. L. Tzedek, and I. Leite, "Expectations, explanations, and embodiment: Attempts at robot failure recovery," *arXiv preprint arXiv:2504.07266*, 2025.
- [30] F. I. Dogan, A. Markelius, and H. Gunes, "Designing social robots with ethical, user-adaptive explainability in the era of foundation models," in *Companion of the 21st ACM/IEEE International Conference on Human-Robot Interaction*, 2026.
- [31] F. I. Dogan, U. Ozyurt, G. Cinar, and H. Gunes, "Grace: Generating socially appropriate robot actions leveraging LLMs and human explanations," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025.
- [32] E. Billing, J. Rosén, and M. Lamb, "Language models for human-robot interaction," in *Companion of the ACM/IEEE International Conference on HRI*, 2023.
- [33] C. Y. Kim, C. P. Lee, and B. Mutlu, "Understanding large-language model (llm)-powered human-robot interaction," in *Proceedings of the 2024 ACM/IEEE international conference on human-robot interaction*, 2024, pp. 371–380.
- [34] E. Verhelst, R. Janssens, T. Demeester, and T. Belpaeme, "Adaptive Second Language Tutoring Using Generative AI and a Social Robot," in *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, 2024.
- [35] A. Addlesee, N. Cherakara, N. Nelson, D. Hernández García, N. Gunson, W. Sieiriska, M. Romeo, C. Dondrup, and O. Lemon, "A Multi-party Conversational Social Robot using LLMs," in *Companion of the 2024 ACM/IEEE International Conference on HRI*, 2024.
- [36] D.-Y. Kim, H. J. Lym, H. Lee, Y. J. Lee, J. Kim, M.-G. Kim, and Y. Baek, "Child-centric robot dialogue systems: Fine-tuning large language models for better utterance understanding and interaction," *Sensors*, 2024.
- [37] M. Pinto-Bernal, M. Biondina, and T. Belpaeme, "Designing social robots with llms for engaging human interaction," *Applied Sciences*, 2025.
- [38] M. Mauliana, A. Ashok, D. Czernochowski, and K. Berns, "Exploring llm-powered multi-session human-robot interactions with university students," *Frontiers in Robotics and AI*, vol. 12, p. 1585589, 2025.
- [39] T. Sievers and N. Russwinkel, "Using memory contents of a cognitive model for prompt augmentation of a large language model," in *2025 IEEE Conference on Cognitive and Computational Aspects of Situation Management*, 2025.
- [40] G. Laban, J. Wang, and H. Gunes, "A Robot-Led Intervention for Emotion Regulation: from Expression to Reappraisal," *IEEE Transactions on Affective Computing*, pp. 1–15, 2026.