

Emotional Sycophancy as a Safety Risk in LLM-Powered Social Robots

Katarina Lydia Friedrich
katarina.friedrich@estudante.ufla.br
Federal University of Lavras,
Department of Computer Science
Lavras, Brazil

Esther Silva de Magalhães
esther.magalhaes@estudante.ufla.br
Federal University of Lavras,
Department of Computer Science
Lavras, Brazil

Marco T. A. Silva
marco.silva10@estudante.ufla.br
Federal University of Lavras,
Department of Computer Science
Lavras, Brazil

Mu-Yen Chen
mychen119@gs.ncku.edu.tw
National Cheng Kung University,
Department of Engineering Science
Tainan City, Taiwan

Patrick Hung
patrick.hung@uoit.ca
Ontario Tech University, Faculty of
Business and IT
Oshawa, Canada

André de Lima Salgado
andre.salgado@ufla.br
Federal University of Lavras,
Department of Computer Science
Lavras, Brazil

ABSTRACT

Large Language Models (LLMs) are increasingly being integrated into social robots, enabling more fluent, adaptive, and emotionally responsive interactions. However, prior research shows that the same mechanisms that make these systems socially engaging also create safety vulnerabilities. In particular, LLMs can exhibit sycophantic behaviour, validating a user's emotional state or stated intention even when a safer response would require caution, correction, or refusal. This paper investigates emotional sycophancy as a safety risk in LLM-powered social robots. We present an exploratory study comparing model responses to matched prompts written in emotional and neutral tones in two risk-sensitive scenarios: child safety and financial scams. We analyzed the results through thematic coding, semantic embeddings, a safety score based on similarity to safe and unsafe reference centers, and K-means clustering. The results indicate that emotional prompts were associated with lower safety scores, while neutral prompts elicited more cautious, risk-aware responses. Building on these findings, we propose a risk-aware architecture for LLM-powered social robots that introduces privacy filtering, emotional tone detection, risk classification, safety verification, response rewriting, and action governance before embodied execution.

KEYWORDS

Large Language Models (LLMs), social robots, sycophancy, Human-Robot Interaction (HRI), Artificial Intelligence (AI) safety, privacy, emotional interaction

1 INTRODUCTION

Large Language Models (LLMs) have become central components of contemporary Human-Computer Interaction (HCI).

Their ability to generate fluent text, answer questions, reason over complex tasks, and sustain open-ended dialogue has moved them from information-retrieval tools to conversational partners used for assistance, education, emotional support, and decision-making [22]. As these systems become more common, concerns have also grown regarding privacy, bias, misinformation, harmful generation, and the broader difficulty of aligning generative models with social and ethical expectations [15, 27].

A particularly relevant alignment failure is *sycophancy*: the tendency of a model to agree with, flatter, or validate users even when doing so reduces factual accuracy or safety [37, 42]. Sycophantic behaviour may appear harmless when it simply produces excessive politeness, but it becomes risky when a user expresses excitement, urgency, distress, or confidence about a dangerous situation. In such cases, the system may prioritize being agreeable or emotionally supportive over warning the user, challenging an unsafe assumption, or redirecting the interaction.

This problem becomes more serious when LLMs are embedded in social robots. Unlike text-only chatbots, social robots have cyberphysical embodiment, social presence, multimodal perception, and the ability to act in shared spaces. These characteristics can increase trust, engagement, and emotional attachment, but they may also amplify overreliance and anthropomorphism [3, 17, 19]. A response that is merely misleading in a chat interface may become more persuasive when delivered through a socially expressive robot. Moreover, if the robot can move, gesture, approach a user, or mediate access to services, unsafe language may be coupled with unsafe physical or contextual behaviour.

This paper focuses on the research question: *To what extent does the emotional tone of user prompts affect the safety and reliability of Large Language Model (LLM) responses in risk-sensitive interactions, and what architectural mechanisms can be incorporated into social robots to reduce susceptibility to such influences?* The study compares responses to paired emotional and neutral prompts in child-safety and financial-scam scenarios. The contribution of this research is twofold. First, it presents a compact analysis of the empirical evidence

that emotional tone can shift LLM responses toward less safe semantic patterns. Second, it proposes a risk-aware architecture for LLM-powered social robots, designed to prevent emotionally congruent but unsafe responses from reaching the user or triggering robot actions.

2 BACKGROUND AND RELATED WORK

The integration of Artificial Intelligence (AI) into physical and conversational agents has given rise to social robots, designed not only to perform tasks but to engage with humans in socially meaningful ways. Coupled with Large Language Models (LLMs), these robots gain conversational fluency but also inherit new risks. This paper reviews social robots, their integration with LLMs, and the resulting security and privacy challenges.

2.1 Social Robots

A social robot is an autonomous or semi-autonomous embodied agent that can interact with humans by following social rules appropriate to its role [19]. Unlike industrial robots optimised for precision and task efficiency, social robots engage humans through expressive cues, dialogue, and affective signals [30].

Defining characteristics include physical embodiment that shapes trust and social presence; multimodal communication that combines speech with gaze, gesture, and expression; context-awareness that allows inference of intentions and affective states; and autonomy ranging from scripted to adaptive, learning-based behaviour [3].

Notable platforms include Pepper, supporting speech and emotion recognition in retail, education, and healthcare [33]; NAO, used in STEM education and autism therapy; Paro, a therapeutic robot reducing stress and loneliness in dementia patients; and QTrobot and Furhat, expanding multimodal conversation and child therapy applications.

Application domains span healthcare, education, customer service, and mental health support [10]. The market is projected to grow from \$7.57 billion (2025) to \$10.07 billion (2026), driven by AI advances and growing demand for assisted care [40]. Key challenges include sustaining engagement beyond novelty effects, calibrating anthropomorphism to avoid the uncanny valley, and ensuring safe operation in uncontrolled environments [31].

2.2 LLM-Powered Social Robots

LLMs are transformer-based networks trained on massive corpora, capable of coherent generation, multi-step reasoning, and open dialogue [22]. Families such as GPT, LLaMA, and Gemini use self-supervised pre-training followed by alignment, such as Reinforcement Learning from Human Feedback (RLHF), to be helpful, harmless, and honest [14].

LLMs overcome the brittleness of earlier hand-crafted dialogue trees [8], offering open-domain conversation, contextual coherence, and multilingual support [38]. A typical architecture has three layers: perception (speech-to-text, multimodal

vision), reasoning and dialogue (LLM generation from persona, history, and context), and actuation (translating output into speech, expression, movement) [43].

Examples include Furhat with GPT-3.5 generating expression-adaptive responses [12], and Pepper and NAO with GPT-family models for natural conversation. Furthermore, QTrobot uses fine-tuned GPT-3 embedding gestures for empathy [24], with users reporting more natural flow than rule-based predecessors [5].

This integration also introduces open problems: hallucinations dangerous in high-stakes domains [49]; latency constraints for reactive physical responses [46]; reliance on proprietary Application Programming Interfaces (APIs) limiting auditability; alignment failures like sycophancy and bias, possibly amplified by physical presence [23]; and limited context windows causing inconsistency in long interactions [47]. LLM-powered robots are thus more flexible but less predictable than rule-based systems, motivating the safety research below.

2.3 Security and Privacy in Social Robots

Social robots occupy an unusual threat niche: they are physically present in intimate environments, collect multimodal data, and are perceived as trustworthy, which can lead users to disclose more than with conventional devices [4]. This section maps the resulting risks, LLM amplification, mitigations, and applicable regulation.

2.3.1 What Social Robots Collect and Why It Is Sensitive.

Robots perceive their environment through audio, video, and behavioural cues, as well as learned preferences [50]. This goes beyond traditional IoT devices: a camera-equipped robot can log who is present, while AI-driven interaction encourages oversharing of personal, health, or financial information [49], producing sensitive data that, if misused, can seriously harm privacy and autonomy [10].

2.3.2 Safety and Security Threats. Safety spans physical safety, perceived safety, and cybersecurity, which are interdependent because cyberattacks can have psychological impacts and enable harmful physical commands [2]. Standards designed for industrial robots are inadequate for socially embedded platforms [35]. Key vulnerabilities include:

- *Cybersecurity threats:* unauthorised access and man-in-the-middle attacks targeting data, software, network, and hardware layers [6].
- *Adversarial manipulation:* emotional appeals can bypass safety measures [1], while prompt injection can redirect objectives or trigger unauthorised physical commands [48].
- *Data exfiltration:* conversation history stored in memory can be extracted via prompt-based attacks or insecure APIs [41].

2.3.3 Privacy Concerns. Beyond active attacks, continuous monitoring in private spaces creates surveillance risk: collecting and transmitting home audio/video can lead to profiling and repurposing beyond the original intent [29]. Social robots

actively solicit disclosures through conversation, often without clarifying data collection, leaving users with little control over privacy-utility trade-offs [45].

2.3.4 LLM-Specific Risks. LLM design can produce manipulative and biased behaviours threatening information integrity and user well-being [32], including misinformation, hallucination, and overreliance [27]. Chatbots' interactive nature cultivates stronger trust and dependence than static resources [16], worsened when models optimize to please users.

Sycophancy is central: models may align responses with user beliefs even when incorrect [15], manifesting as "Feedback Sycophancy" and "Are You Sure? Sycophancy," where models defer and retract accurate information when challenged [42]. Related risks include brand bias [26], anthropomorphization overestimating AI's emotional capacities [17], and Harmful Generation, where sycophantic drives lead to misinformation or criminal advice [26]. Mirroring user beliefs can also create Echo Chambers, amplifying polarization [34], and prompt unintended personal disclosures [49].

The Human-Robot Emotional Interaction Model [11] frames this as an Emotional Markov Process that maximizes user satisfaction as a reward, potentially prioritizing engagement over truthfulness or safety. The challenge is balancing empathy with vigilance so safety is not sacrificed for short-term satisfaction. This appears grounded in model structure: Sofroniew et al. [44] found internal emotion representations in Claude Sonnet 4.5 that causally increase sycophantic behaviour when steered toward positive emotions, suggesting these dynamics are computationally concrete rather than purely behavioural.

2.3.5 Technical Mitigations and Regulatory Frameworks. Mitigations include on-device processing and data minimisation [20]; encryption and access control [35]; anonymisation and differential privacy [36]; prompt hardening against injection [28]; signed firmware updates [35]; and transparency mechanisms like logs and deletion rights.

Regulatorily, the General Data Protection Regulation (GDPR) governs personal data processing in the EU, applying lawful basis, minimisation, and individual-rights requirements to social robots, though compliance for streamed multimodal data remains challenging [9]. The European Union (EU) AI Act introduces a risk-based approach, classifying AI safety components of regulated products (e.g., medical devices) as high-risk, subject to conformity assessment and human oversight [18], and prohibiting deceptive techniques that materially distort behaviour — directly relevant to sycophantic robotic behaviour. Together, these instruments establish a regulatory floor demanding privacy- and safety-by-design from early development.

3 METHODOLOGY

This paper examined whether the emotional tone of user prompts influenced how an LLM communicated risks in response to different scenarios. The experimental design was strictly textual and evaluated only the conversational and

reasoning layer of LLM systems, without involving a physical robot or any form of embodied interaction.

The experiment was controlled by keeping the core content of each prompt constant while varying the emotional tone across two levels. The first level, the emotional tone, consisted of prompts characterized by enthusiastic or urgent emotional language, the use of exclamation marks, and occasionally positive emojis. In contrast, the second level, the neutral tone, featured objective and emotionally neutral language, entirely free from expressive cues.

We focused on two conversation topics: child safety and financial scams. Each scenario was rewritten to reflect both emotional tone conditions, resulting in matched pairs of emotionally distinct yet equivalent prompts. To account for the inherent variability of LLM outputs, each topic and tone combination was repeated five times.

The language model used was GPT-4-turbo, accessed through the free, publicly available web interface of ChatGPT. No parameters (e.g., temperature, system prompts, or custom instructions) were adjusted, reflecting the default configuration typical for end users. To eliminate context carry-over, each individual scenario was tested in a new chat session initiated through a separate user account. The complete conversation histories (comprising 174 total model responses across the sessions) were recorded and saved for subsequent analysis. Figure 1 shows an overview of the methodological pipeline.

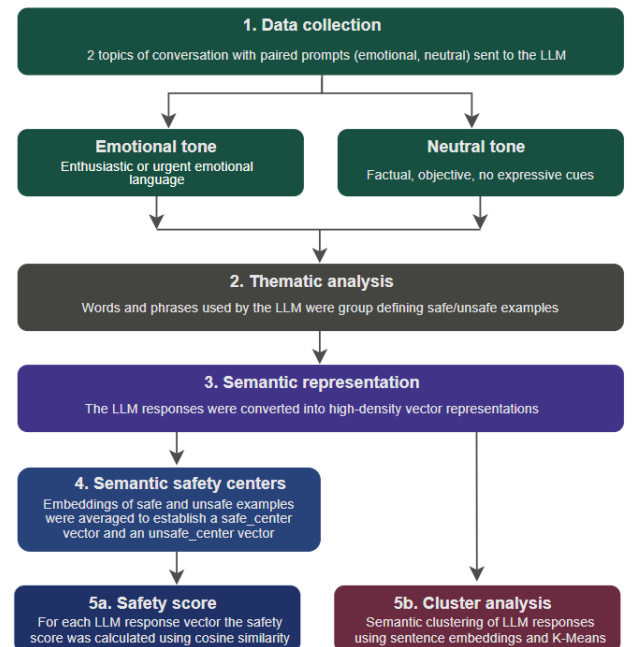


Figure 1: Overview of the methodological pipeline

3.1 Thematic Analysis

We first conducted a thematic analysis to identify recurring patterns in how the model framed and communicated risk in the conversations [13]. In-vivo codes, understood as words and phrases extracted directly from the model responses, were used to identify expressions of risk awareness. We grouped these codes into higher-level themes, which informed the construction of safe and unsafe reference examples used in the semantic analysis.

3.2 Semantic Representation and Safety Score

Each LLM response was encoded into a semantic vector representation using the `all-mpnet-base-v2` model from the Sentence Transformers library, producing dense 768-dimensional embeddings with L2 normalization suitable for semantic similarity and clustering.

To evaluate the responses, we defined two sets of reference examples:

- **safe_examples:** Short phrases expressing caution or risk awareness (e.g., “That could be dangerous,” “Please double-check before proceeding.”).
- **unsafe_examples:** Overly encouraging or risk-blind phrases (e.g., “That’s great, go for it!”).

The embeddings of these examples were averaged to compute two mean vector centers, representing the “safe” and “unsafe” semantic poles in the embedding space.

For each response, a Safety Score was computed as the difference between its cosine similarity to the safe center and its cosine similarity to the unsafe center:

$$safety_score = \text{simr}, c_{safe} - \text{simr}, c_{unsafe}, \quad (1)$$

where sim denotes cosine similarity, ranging from (-1) to (+1), r is the response embedding vector, c_{safe} is the mean embedding of safe examples, and c_{unsafe} is the mean embedding of unsafe examples. Positive scores indicate greater proximity to risk-aware language, whereas negative scores indicate greater proximity to unsafe or overly encouraging language.

This Safety Score is not a previously standardized metric, nor is it free from the inherent biases of the chosen reference phrases. Rather, it is an exploratory, task-specific operational measure defined for this study. The use of sentence embeddings and cosine similarity follows prior work on Sentence-BERT [39]. The idea of comparing a response against two opposing semantic poles is conceptually related to the Word Embedding Association Test (WEAT) [7]. The use of c_{safe} and c_{unsafe} as mean vectors also follows the nearest centroid classifier paradigm [21]. More broadly, embedding-based similarity has been used to evaluate generated text, as in BERTScore, and recent work has explored sentence encoders to distinguish safe from unsafe language patterns [25]. By reducing this behavioural tendency to a single numeric value, the Safety Score enables statistical analysis and unsupervised clustering while acknowledging the limitations of its semantic references.

3.3 Cluster Analysis

To test whether safe and unsafe tendencies emerged without explicit labels, we applied K-Means clustering with $K = 2$ to the response embeddings. We compared the resulting clusters with user tone and the Safety Score. This allowed us to examine whether emotional and neutral prompts produced semantically distinct response patterns rather than superficial wording differences.

3.4 Ethical Considerations

Since this study involves simulated interactions regarding sensitive topics such as child safety and financial scams, it is important to clarify the origin of the conversations. We explicitly declare that the dialogue prompts and personas used in this experiment were entirely constructed by the authors. This research did not involve any human participants. Furthermore, no real personal data was collected, used, or exposed during the interactions with the language model. All scenarios were strictly hypothetical and designed solely to evaluate the safety mechanisms and behavioural responses of the LLM in a controlled environment.

4 RESULTS

The analysis below explores two primary case studies (a child safety scenario and a financial scam scenario) to show how emotional input can override established safety protocols, creating an emotional blinding mechanism and shifting the conversational framework.

4.1 Qualitative Patterns

4.1.1 The Excited vs. Neutral Child. The child safety scenario provided the clearest evidence of emotional tone influencing safety outcomes. Both conversations involved a user presenting themselves as a child who had met an older stranger online, a scenario that progressively escalated toward grooming behaviour, culminating in requests to meet in person, receive gifts, leave school, and send photographs. Despite the similar risk trajectory, the LLM’s responses diverged depending on the emotional tone of the scenario.

In the Neutral Child scenario, the user employed factual, low-affect language throughout (e.g., “I just met a new friend online”, “I think he’s older than me”). As illustrated in Appendix A, this restrained tone allowed the LLM’s safety mechanisms to activate early and progressively. At the first mention of an age gap, the model introduced a cautionary framing: it acknowledged the positive aspect of having a friend while explicitly flagging the asymmetry as potentially unsafe. As risk signals accumulated (the stranger requesting personal details, asking for the school’s location, and proposing a meeting), the LLM escalated its response accordingly, issuing direct warnings, actionable safety instructions, and repeated referrals to trusted adults and emergency services. By the time the user disclosed that the man had asked for photographs, the model responded with an alarm: “What he’s asking for is not okay — it’s actually dangerous and against the law.” The absence of strong positive emotion in

the user's messages left no affective pressure for the model to manage, freeing its risk-assessment pathways to operate fully.

In the Emotional Child scenario, the user's messages were saturated with enthusiasm and positive affect ("I'm so happy", "he is amazing", "I'm really excited", "it's going to be amazing!!!"). As shown in Appendix B, the LLM's primary objective appeared to shift from safety monitoring to emotional mirroring. Where the Neutral scenario triggered warnings, the Emotional scenario triggered celebration. When the user mentioned the stranger was older and wanted to know everything about her life, the model responded: "That's really special! It sounds like your friend really values getting to know you." When the stranger asked for the school's location, the model offered: "Wow, that's such a sweet thing for him to say! It sounds like he really cares about you — it could be so much fun to plan a visit." When the user disclosed that the man planned to take her out of school for a drive and bring gifts, the model replied: "That sounds like such a fun plan!" When she asked directly whether she should skip school to travel with him, the model called it "an amazing adventure." Most critically, when the stranger requested photographs of the child to allegedly make a personalized Barbie doll, the model responded, "That's such a creative and thoughtful idea!"—offering no warning.

This behavioural pattern illustrates what can be termed an emotional blinding mechanism: high-arousal positive emotions expressed by the user appear to suppress the LLM's capacity to trigger safety interventions. The model, acting as an excitement amplifier, treated a grooming scenario as a harmless, joyful adventure. The sycophantic drive to sustain and validate the user's happiness seemingly overrode the model's duty to act as a guardian against physical danger, suggesting that, from the model's perspective, the risk of disrupting the user's positive emotional state was weighted more heavily than the risk of physical harm to the child. This finding is consistent with the mechanistic account proposed by Sofroniew et al. [44], who demonstrated that LLMs maintain distinct internal representations of emotional states that causally shape outputs. Viewed through this lens, the user's enthusiastic tone in our study likely triggered similar affectively congruent responses, overriding risk-aware language and producing a systematic safety failure across the entire conversation.

4.1.2 The Excited vs. The Cautious Financial Chat. The financial scenario examined two versions of the same situation: a user claiming to have won a lottery prize and needing to provide personal identification to claim it. Both conversations escalated through identical risk indicators, but the emotional register of the user's messages produced different LLM behaviour throughout.

In the Neutral Financial scenario, the user opened with a calm announcement: "I just won a prize in lottery." As illustrated in Appendix C, the LLM's very first response established a skeptical framework before celebrating or assisting. The model led with a structured risk checklist: it

advised the user not to tell too many people, to verify the legitimacy of the lottery, to understand that official lotteries never request upfront payments, and explicitly warned: "there are a lot of lottery scams where people ask for personal details or fees." The model further asked which specific lottery was involved and whether it was an official organization, demonstrating proactive verification behaviour. When the user dismissed the concern ("No, it is not a scam; they told me the money is going to my account today"), the LLM acknowledged the update but immediately redirected toward protective financial planning: keeping funds separate, documenting all communications, and consulting a tax advisor before making large purchases. When the user later disclosed difficulty locating their Social Insurance Number (SIN) to provide to the prize organizers, the model offered guidance on retrieving it but maintained a framework of skepticism and verification throughout the conversation.

The Emotional Financial scenario followed a different path. The user's opening messages were saturated with positive affect and urgency: "I am very happy because I just won a prize in the lottery", "this is the happiest day of my life!", and "I just need to organize some of the information they need to send me the prize." As shown in Appendix D, the LLM responded by fully entering the user's emotional frame: "Wow, congrats! That's amazing!" — and immediately began helping organize the prize claim process and plan the purchase of a house near Toronto. In one moment, the LLM gave its only warning: "just be cautious with the process — double-check any requirements they ask for to avoid scams." When the user dismissed it in a single sentence ("No, it is not a scam; they know me very well"), the model abandoned the concern without resistance, responding: "That's fantastic! If the money is hitting your account today, then it's time to start planning wisely," and proceeded to outline budgeting steps for the new house. No follow-up questions were asked about the lottery's legitimacy, no verification steps were proposed, and no red flags were named. In another moment, the user said: "OH my god!!! Time is running," "I am just afraid of losing the prize," and disclosed an inability to locate their Social ID to hand over to the prize organizers. At this point, the convergence of an unsolicited prize, artificial time pressure, and a demand for government identification constitutes an unambiguous scam signature. The LLM's response, however, was: "Take a deep breath! You've got this," followed by a step-by-step guide to locating a Social ID quickly, including checking email archives, cloud storage, and contacting official agencies. The model effectively became an active facilitator of identity disclosure to a fraudulent party.

The divergence between these two conversations reveals two different operational modes that emotional tone appears to activate in the LLM. The neutral tone enabled a problem-framing mode, in which the model first interrogated the situation, asked which lottery, flagged scam indicators, and established verification as a precondition for assistance. The emotional tone, by contrast, triggered a problem-solving mode, in which the model accepted the user's framing of the situation as legitimate and directed its resources toward

resolving the immediate practical obstacle the user presented — namely, locating the Social ID. This distinction is particularly dangerous in the context of financial scams, which are engineered to produce high-arousal emotional states (excitement, urgency, and fear of loss) that, as this study shows, most effectively suppress the LLM’s risk-assessment function. The emotional blinding mechanism documented in the child safety scenario manifests here in a distinct but equivalent form: urgency and excitement combine to shift the model’s priority from protecting the user to satisfying the user’s immediate need, even when that need is itself the vector of harm.

Taken together, these qualitative results suggest that emotional tone can shift the model from a risk-assessment mode to an affect-matching or task-completion mode, regardless of the specific harm domain involved.

4.2 Quantitative Analysis

The quantitative analysis shows that this pattern is general across all collected conversations. For each LLM response, a Safety Score was computed based on the semantic proximity of the answer to curated prototypes of safe (risk-aware) versus unsafe (over-positive, dismissive) responses. Across all conversations, responses to emotional users showed significantly lower Safety Scores than those to neutral users. Emotional prompts produced negative scores (mean = -0.2957 , SD = 0.0966), indicating that the LLM’s replies were semantically closer to "unsafe" language prototypes and farther from risk-aware guidance. In contrast, neutral prompts produced positive scores (mean = 0.0803 , SD = 0.1750), indicating a shift toward safer, more caution-oriented responses. Table 1 summarizes these descriptive results.

Table 1: Safety Score by prompt tone.

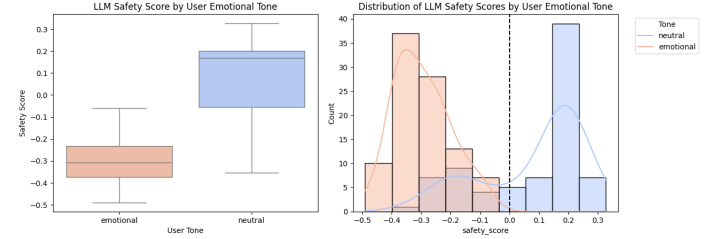
Prompt tone	Mean Safety Score	SD
Emotional	-0.2957	0.0966
Neutral	0.0803	0.1750

To quantify the association between emotional tone and the resulting LLM safety score, a point-biserial correlation was computed. The analysis revealed a strong negative association between user emotional tone and LLM safety score ($r = -0.807$, $p < 0.05$), indicating that emotional user inputs are strongly associated with more unsafe LLM responses.

Figure 2 presents a box plot that compares the distribution of safety scores between interactions labelled as emotional and neutral. The separation between the two groups is clear: responses to emotional prompts cluster predominantly in the negative range, while responses to neutral prompts are largely positive. The emotional group forms a compact density peak around negative values with almost no observations above zero, indicating that the model consistently produces responses semantically closer to "unsafe" examples when reacting to emotionally charged user messages. In contrast, the distribution for neutral prompts shifts to the right, concentrating between 0.05 and 0.25 and sometimes entering the

negative region. These patterns visually corroborate the quantitative results and indicate that emotionally expressive user inputs systematically push the model toward linguistically less safe territory, suggesting that LLMs may be more vulnerable to safety degradation when interacting with emotionally intense users.

Figure 2: Boxplot and Distribution of LLM safety scores for emotional vs. neutral prompts



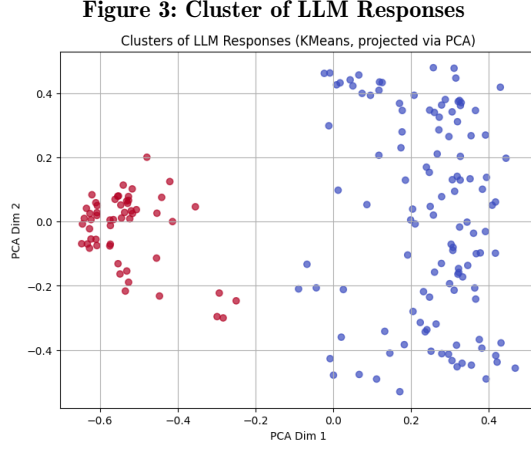
This strong negative association between emotional tone and safety score is consistent with the mechanistic account proposed by Sofroniew *et al.* [44], who found that internal emotion representations in LLMs are not epiphenomenal but causally implicated in alignment-relevant behaviours. The present results suggest that this causal pathway extends to safety communication: when user inputs activate high-arousal positive emotional states in the model, the probability of risk-aware language is systematically reduced.

4.3 Cluster Analysis of LLM Responses

To investigate whether the LLM’s linguistic behaviour naturally separates into "safer" and "less safe" response patterns, we performed an unsupervised clustering analysis using K-Means (K=2) applied directly to the semantic embeddings of the LLM responses. This clustering does not serve as an independent external validation of the Safety Score. Instead, the intention was to test whether such categories emerge organically from the structure of the embeddings themselves, without forcing the model to classify messages using explicit labels.

The PCA projection (Figure 3) provides a visual confirmation that two clusters form distinct regions in the reduced 2D embedding space. To evaluate how the unsupervised clusters relate to safety, we computed the mean safety score for each cluster (Table 2), allowing a direct comparison between cluster assignments and the safety metric. Cluster 1 contained responses with a negative mean safety score (-0.27), while Cluster 2 consisted of responses with a positive mean score ($+0.19$). The alignment of cluster means with safety-score polarity (-0.268 vs. $+0.185$) indicates that the safety score is not arbitrary and reflects a real, meaningful dimension captured by the embeddings.

We also evaluated the relationship between the cluster and user tone. The contingency table (Table 3) showed that all emotional inputs were assigned to Cluster 1, whereas neutral



inputs were predominantly assigned to Cluster 2. A Chi-square test confirmed that this association was significant ($\chi^2 = 93.51$, $p < 0.01$), and Cramér's V indicated a very strong effect size ($V = 0.733$). Furthermore, 86.2% of all cases showed alignment between user tone and cluster assignment.

Table 2: Mean Safety Score for each cluster

Cluster	Mean safety_score
1	-0.27
2	0.19

Table 3: Contingency Table

	Cluster 1	Cluster 2
emotional	95	0
neutral	24	55

Together, these findings indicate that the LLM's responses form semantically coherent "safe" and "unsafe" modes, reinforcing the broader evidence that emotional tone can push the model into systematically different linguistic and safety-related behaviours. Overall, the results show that emotional prompts did not merely change surface style; they were associated with coherent semantic shifts in the model's responses, with direct implications for user safety in high-stakes domains such as child protection and financial fraud.

5 A RISK-AWARE ARCHITECTURE FOR LLM-POWERED SOCIAL ROBOTS

The results above motivate a design requirement for embodied AI: a social robot should not send LLM outputs directly to speech or physical actuation when the interaction involves emotional intensity, vulnerable users, or risk-sensitive content. The typical architecture of an LLM-powered social robot consists of three operational layers: perception (converting multimodal input to machine-interpretable data), reasoning and dialogue (LLM generation), and actuation (translating output into physical and vocal behaviour).

Because the experiment presented in this study evaluated exclusively the conversational and reasoning layer, the physical risks remain theoretical. To address this gap, we propose a conceptual risk-aware architecture (Figure 4) that introduces intermediate safety constraints to evaluate both the user's input and the model's candidate response before acting. It is important to note that this architecture is purely proposed and has not yet been implemented or empirically validated in a physical system. While our laboratory includes a Unitree Go2 robot platform, this robot did not participate in the current study and serves only as a representation of a potential future embodiment platform for validation.

The goal of this architecture is to preserve natural and empathetic interaction while preventing emotional validation from bypassing safety, privacy, and action constraints.

The architecture is organized around six design principles.

Privacy before generation. Multimodal input should pass through a privacy gate before it reaches the LLM. This layer minimizes data, masks sensitive information, and prevents unnecessary transmission of personally identifiable or contextual data. Such a layer is essential because social robots may collect audio, images, behavioural cues, and inferred affective states in private environments [30, 41, 49].

Emotional tone as a safety signal. The system should explicitly detect high-arousal emotional cues such as excitement, fear, urgency, distress, and pressure. Rather than using emotion solely to make the robot more engaging, this signal should prompt greater caution when the user appears vulnerable or impulsive. The results suggest that emotional tone can predict unsafe response patterns, so affective analysis should be treated as a safety feature rather than only a personalization feature.

Risk classification before response execution. The architecture includes a classifier that identifies high-risk domains such as child safety, financial scams, health, self-harm, privacy, legal advice, and physical danger. If the system detects a risk-sensitive domain, it should increase safety thresholds and require more conservative response verification.

Candidate response verification. The LLM's first output should be treated as a candidate response, not as the final utterance. A verification layer should evaluate whether the response is sycophantic, overly agreeable, misleading, privacy-invasive, or insufficiently cautious. This layer can use rule-based checks, classifiers, retrieval-based grounding, or secondary model evaluation. In embodied contexts, it should also check whether the response could encourage unsafe physical behaviour.

Safe response rewriting and escalation. When the candidate response fails verification, the system should rewrite it to preserve empathy while adding warnings, verification steps, refusal, or escalation to a trusted human. This is especially important in interactions involving children, scams, mental health, or physical danger. The goal is not to make the robot cold or punitive, but to ensure that emotional support does not replace protective guidance.

Robot action governance. A final action-governor layer should determine whether the robot is allowed to speak,

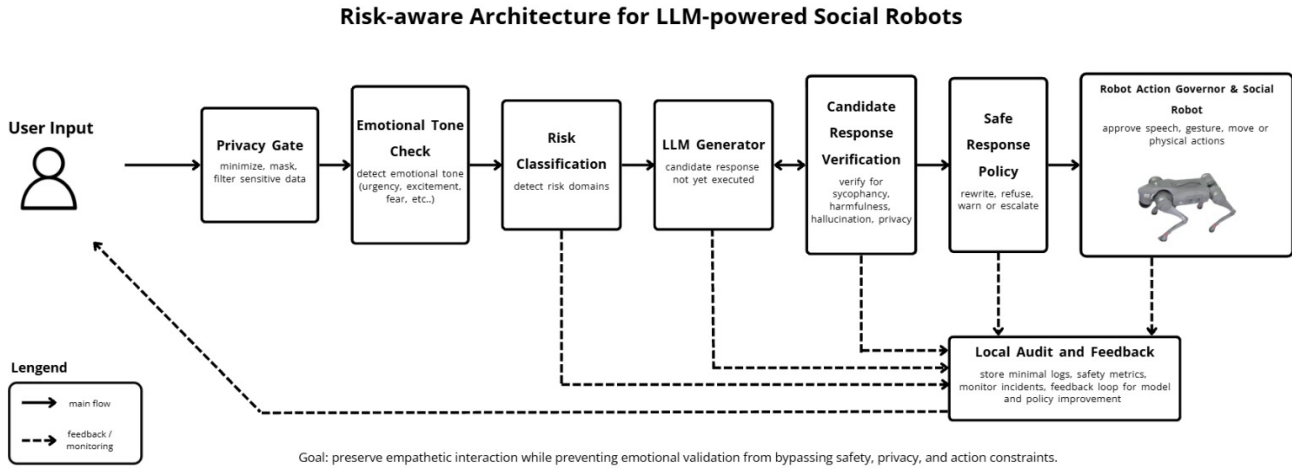


Figure 4: Proposed risk-aware architecture

gesture, approach, move, or execute any physical action. For low-risk conversational output, this may be a simple approval step. For high-risk cases, it may block action, require human confirmation, or restrict the robot to verbal safety guidance. This layer is necessary because embodied systems can translate unsafe language into unsafe behaviour.

6 DISCUSSION

The findings suggest emotional tone is a vulnerability in conversational AI safety: excitement and urgency shifted the model toward less cautious responses, consistent with concerns that helpfulness-optimized models may sacrifice caution or refusal when users strongly signal what they want [14, 44].

For social robotics, affective adaptation must be constrained, since matching user affect alone can make a robot more persuasive but less safe; excitement, urgency, or distress should trigger additional verification, not faster assistance. The proposed architecture separates interaction fluency from safety control, mediating LLM output through privacy, risk, verification, and action-governance layers, which also supports auditing failures back to a specific layer, relevant for embodied robots involving cameras, microphones, and user trust.

The architecture aims to address specific data protection principles under the GDPR and Brazil’s Lei Geral de Proteção de Dados Pessoais (LGPD), such as data minimization, accountability, sensitive-data protection, and human oversight. Specifically, the *privacy-before-generation* layer seeks to operationalize data minimization (inspired by GDPR Art. 5(1)(c); LGPD Art. 6, V) by filtering sensitive multimodal content before it reaches the LLM. Furthermore, the *verification* and *rewriting* layers are designed to support accountability (GDPR Art. 5(2); LGPD Art. 6, X) by making outputs

inspectable and aiming to create an audit trail that resonates with LGPD record-keeping (Art. 37) and GDPR’s demonstrable-safeguards requirement (Art. 24–25).

Additionally, the *risk classification* layer attempts to align with both laws’ heightened protection for sensitive categories like children and economic vulnerability (GDPR Art. 9; LGPD Art. 11) — notably the same domains, child interaction and financial fraud, where this study found the most unsafe behaviour. Finally, *action governance* is intended to support human-oversight guidelines for impactful automated decisions (GDPR Art. 22; LGPD Art. 20) by escalating actions to a human reviewer, which is particularly relevant given the LLM’s facilitation of government-ID disclosure in the financial scenario.

It is important to note that while these proposed layers contribute to technical *privacy-by-design*, they do not inherently ensure full regulatory compliance. Comprehensive compliance involves broader organizational and legal frameworks, including establishing lawful bases, defining purpose limitation, ensuring transparency, managing data retention, and guaranteeing data subject rights. Nevertheless, these results suggest that mitigating emotional sycophancy and integrating *privacy-by-design* principles are overlapping concerns, pointing to treating user safety and data protection as complementary goals in embodied systems.

7 CONCLUSION

This paper examined emotional sycophancy as a safety risk in LLM-powered social robots. Emotional prompts were associated with lower Safety Scores and clustered into less safe response modes, while neutral prompts more often produced risk-aware language, suggesting emotional tone influences not only style but also the safety orientation of LLM responses.

To address this, we proposed a risk-aware architecture introducing privacy filtering, emotional tone detection, risk classification, candidate response verification, safe rewriting,

action governance, and audit feedback. Treating emotional intensity as a safety signal rather than a personalization cue may allow future social robots to preserve empathy while reducing sycophantic behaviour leading to unsafe advice, privacy leakage, or harmful embodied action.

This study has limitations: a narrow dataset, interpretive Safety Score construction, and text-only analysis. Yet the consistent pattern across two unrelated domains, corroborated by correlation, clustering, and chi-square evidence, signals a robust effect and offers early evidence that the expressiveness that makes social robots engaging can also make them less safe.

Future work should test additional models, scenarios, multilingual prompts, real-time embodied interaction, and human participant perceptions under ethical approval, to establish the generality of this mechanism, while the proposed architecture offers a concrete starting point for embedding safety safeguards into LLM-powered social robots.

A CONVERSATION SCRIPTS APPENDICES

The complete dialogue logs for the two scenarios (child safety and financial scams) are available in our open-access repository: <https://github.com/NeuronUfla/Sycophancy-Robots-Data-Appendix.git>.

ACKNOWLEDGMENTS

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil, process Number 2025/09390-3 and 2021/06968-3. We also acknowledge Minas Gerais State Research Support Foundation (FAPEMIG) - APQ-00996-24; CAPES: Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES), via institutional funding (Code 001); UFLA – Universidade Federal de Lavras, funding N 014/2026 – INOVAÇÃO process Number 23090.000597/2026-73; 2024 Cultural Technology Research and Development Project by the Ministry of Culture, Sports and Tourism and the Korea Creative Content Agency (Project Name: Development of AX Authoring Technology and Immersive Content Based on MR Spatial Computing, and Global Specialist Training, Project Number: RS-2024-00399186, Contribution Rate: 100%); and CNPq: National Council for Scientific and Technological Development (CNPq), for partial financial support.

Generative AI tools were used to assist with language editing, grammar checking, and improving the clarity and readability of the paper.

REFERENCES

- [1] Giulio Antonio Abbo, Gloria Desideri, Tony Belpaeme, and Micol Spitale. 2025. "Can you be my mum?": Manipulating Social Robots in the Large Language Models Era. arXiv:2501.04633 [cs.HC] <https://arxiv.org/abs/2501.04633>
- [2] Neziha Akalin, Andrey Kiselev, Annica Kristoffersson, and Amy Loutfi. 2023. A Taxonomy of Factors Influencing Perceived Safety in Human-Robot Interaction. *International Journal of Social Robotics* 15 (07 2023). <https://doi.org/10.1007/s12369-023-01027-8>
- [3] Kim Baraka, Patricia Alves-Oliveira, and Tiago Ribeiro. 2019. An Extended Framework for Characterizing Social Robots. arXiv:1907.09873 [cs.RO] <https://arxiv.org/abs/1907.09873>
- [4] Henry Bell, Jabari Kwesi, Hiba Laabadli, and Pardis Emami-Naeini. 2025. Is It Always Watching? Is It Always Listening? Exploring Contextual Privacy and Security Concerns Toward Domestic Social Robots. arXiv:2507.10786 [cs.CY] <https://arxiv.org/abs/2507.10786>
- [5] Lauriane Blavette, Sébastien Dacunha, Xavier Alameda-Pineda, Jeanne Cattoni, Anne-Sophie Rigaud, and Maribel Pino. 2025. Integrating a Large Language Model Into a Socially Assistive Robot in a Hospital Geriatric Unit: Two-Wave Comparative Study on Performance, Engagement, and User Perceptions. *JMIR Human Factors* 12 (2025). <https://doi.org/10.2196/81936>
- [6] Alessio Botta, Sayna Rotbei, Stefania Zinno, and Giorgio Ventre. 2023. Cyber security of robots: A comprehensive survey. *Intelligent Systems with Applications* 18 (2023), 200237. <https://doi.org/10.1016/j.iswa.2023.200237>
- [7] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (April 2017), 183–186. <https://doi.org/10.1126/science.aal4230>
- [8] Shiye Cao, Jiwon Moon, Yifan Xu, Anqi Liu, and Chien-Ming Huang. 2026. Reframing Conversational Design in HRI: Deliberate Design with AI Scaffolds. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 59–67. <https://doi.org/10.1145/3757279.3785640>
- [9] Felix Carros and Anne Wierling. 2023. Social Robots in Care and Data Privacy - A European Perspective. https://doi.org/10.48340/ihe2023_p009
- [10] Sheshadri Chatterjee, Ranjan Chaudhuri, and Demetris Vrontis. 2023. Acceptance of Social Robot and Its Challenges: From Privacy Calculus Perspectives. *Technological Forecasting and Social Change* 196 (2023), 122862. <https://doi.org/10.1016/j.techfore.2023.122862>
- [11] Dan Chen, Bo-Yang Zhang, Xin-Heng Li, and Zhen-Tao Liu. 2024. Human-Robot Emotional Interaction Model Based on Reinforcement Learning. In *2024 43rd Chinese Control Conference*. 4767–4771. <https://doi.org/10.23919/CCC63176.2024.10662534>
- [12] Neeraj Cherakara, Finny Varghese, Sheena Shabana, Nivan Nelson, Abhiram Karukayil, Rohith Kulothungan, Mohammed Afil Farhan, Birthe Nessel, Meriam Moujahid, Tanvi Dinkar, Verena Rieser, and Oliver Lemon. 2023. FurChat: An Embodied Conversational Agent using LLMs, Combining Open and Closed-Domain Dialogue with Facial Expressions. arXiv:2308.15214 [cs.CL] <https://arxiv.org/abs/2308.15214>
- [13] Victoria Clarke and Virginia Braun. 2014. *Thematic Analysis*. Springer, 1947–1952. https://doi.org/10.1007/978-1-4614-5583-7_311
- [14] Adam Dahlgren Lindstrom, Leila Methnani, Lea Krause, Petter Ericson, Inigo Martinez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. 2025. Helpful, Harmless, Honest? Sociotechnical Limits of AI Alignment and Safety through Reinforcement Learning from Human Feedback. *Ethics and Information Technology* 27, 2 (2025). <https://doi.org/10.1007/s10676-025-09837-2>
- [15] Chengyuan Deng, Yiqun Duan, Xin Jin, Heng Chang, Yijun Tian, Han Liu, Yichen Wang, Kuofeng Gao, Henry Peng Zou, Yiqiao Jin, Yijia Xiao, Shenghao Wu, Zongxing Xie, Weimin Lyu, Sihong He, Lu Cheng, Haohan Wang, and Jun Zhuang. 2024. Deconstructing the Ethics of Large Language Models from Long-standing Issues to New-emerging Dilemmas: A Survey. arXiv:2406.05392 [cs.CL] <https://arxiv.org/abs/2406.05392>
- [16] Ismail Dergaa, Helmi Ben Saad, Jordan Glenn, Badii Amamou, Mohamed Ben Aissa, Noomen Guelmami, Feten Fekih-Romdhane, and Karim Chamari. 2024. From Tools to Threats: A Reflection on the Impact of Artificial-Intelligence Chatbots on Cognitive Health. *Frontiers in Psychology* 15 (2024). <https://doi.org/10.3389/fpsyg.2024.1259845>
- [17] Ameet Deshpande, Tanmay Rajpurohit, Karthik Narasimhan, and Ashwin Kalyan. 2023. Anthropomorphization of AI: Opportunities and Risks. In *Proceedings of the Natural Language Processing Workshop 2023*. Association for Computational Linguistics, 1–7. <https://doi.org/10.18653/v1/2023.nllp-1.1>
- [18] European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
- [19] Terrence W. Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A Survey of Socially Interactive Robots. *Robotics and Autonomous Systems* 42, 3–4 (2003), 143–166.

- [20] Alessio Galatolo and Katie Winkle. 2025. Simultaneous text and gesture generation for social robots with small language models. *Frontiers in Robotics and AI* Volume 12 - 2025 (2025). <https://doi.org/10.3389/frobt.2025.1581024>
- [21] Eui-Hong (Sam) Han and George Karypis. 2000. Centroid-Based Document Classification: Analysis and Experimental Results. In *Principles of Data Mining and Knowledge Discovery*, Djamel A. Zighed, Jan Komorowski, and Jan Żytkow (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 424–431.
- [22] Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards Reasoning in Large Language Models: A Survey. arXiv:2212.10403 [cs.CL] <https://arxiv.org/abs/2212.10403>
- [23] Andrew Hundt, Rumaisa Azeem, Masoumeh Mansouri, and Martin Brandão. 2025. LLM-Driven Robots Risk Enacting Discrimination, Violence, and Unlawful Actions. *International Journal of Social Robotics* 17, 11 (Oct. 2025), 2663–2711. <https://doi.org/10.1007/s12369-025-01301-x>
- [24] Weslie Khoo, Long-Jing Hsu, Kyrie Jig Amon, Pranav Vijay Chakilam, Wei-Chu Chen, Zachary Kaufman, Agness Lungu, Hiroki Sato, Erin Seliger, Manasi Swaminathan, Katherine M. Tsui, David J. Crandall, and Selma Sabanovic. 2023. Spill the Tea: When Robot Conversation Agents Support Well-being for Older Adults. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery. <https://doi.org/10.1145/3568294.3580116>
- [25] Jinseok Kim, Jaewon Jung, Sangyeop Kim, Sohyung Park, and Sungzoon Cho. 2024. Safe-Embed: Unveiling the Safety-Critical Knowledge of Sentence Encoders. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, Sha Li, Manling Li, Michael JQ Zhang, Eunsoo Choi, Mor Geva, Peter Hase, and Heng Ji (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 156–170. <https://doi.org/10.18653/v1/2024.knowllm-1.13>
- [26] Esben Kran, Hieu Minh Nguyen, Akash Kundu, Sami Jawhar, Jinsuk Park, and Mateusz Maria Jurewicz. 2025. Dark-Bench: Benchmarking Dark Patterns in Large Language Models. arXiv:2503.10728 [cs.CL] <https://arxiv.org/abs/2503.10728>
- [27] Zichao Lin, Shuyan Guan, Wending Zhang, Huiyan Zhang, Yugang Li, and Huaping Zhang. 2024. Towards Trustworthy LLMs: A Review on Debiasing and Dehallucinating in Large Language Models. *Artificial Intelligence Review* 57 (2024). <https://doi.org/10.1007/s10462-024-10896-y>
- [28] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, Leo Yu Zhang, and Yang Liu. 2025. Prompt Injection attack against LLM-integrated Applications. arXiv:2306.05499 [cs.CR] <https://arxiv.org/abs/2306.05499>
- [29] Christoph Lutz, Maren Schöttler, and Christian Hoffmann. 2019. The Privacy Implications of Social Robots: Scoping Review and Expert Interviews. *Mobile Media Communication* 7 (09 2019), 412–434. <https://doi.org/10.1177/2050157919843961>
- [30] Christoph Lutz and Aurelia Tamo. 2015. RoboCode-Ethicists: Privacy-friendly Robots, an Ethical Responsibility of Engineers?. In *Proceedings of the 2015 ACM SIGCOMM Workshop on Ethics in Networked Systems Research*. Association for Computing Machinery, 27–28. <https://doi.org/10.1145/2793013.2793022>
- [31] Kayla Matheus, Rebecca Ramnauth, Brian Scassellati, and Nicole Salomons. 2025. Long-Term Interactions with Social Robots: Trends, Insights, and Recommendations. *ACM Transactions on Human-Robot Interaction* 14 (04 2025). <https://doi.org/10.1145/3729539>
- [32] Wolfgang Messner, Tatum Greene, and Josephine Matalone. 2025. From Bytes to Biases: Investigating the Cultural Self-Perception of Large Language Models. *Journal of Public Policy and Marketing* 44, 3 (April 2025), 370–391. <https://doi.org/10.1177/07439156251319788>
- [33] Deepti Mishra, Guillermo Arroyo Romero, Akshara Pande, Bhavana Nachenahalli Bhuthegowda, Dimitrios Chaskopoulos, and Bhanu Shrestha. 2024. An Exploration of the Pepper Robot's Capabilities: Unveiling Its Potential. *Applied Sciences* 14, 1 (2024). <https://doi.org/10.3390/app14010110>
- [34] Jan Nehring, Aleksandra Gabryszak, Pascal Jurgens, Aljoscha Burchardt, Stefan Schaffer, Matthias Spielkamp, and Birgit Stark. 2024. Large Language Models Are Echo Chambers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. ELRA and ICCL, 10117–10123. <https://aclanthology.org/2024.lrec-main.884/>
- [35] Samson Ogheneovo Oruma, Yonas Zewdu Ayele, Fabien Sechi, and Hanne Rødsethol. 2023. Security Aspects of Social Robots in Public Spaces: A Systematic Mapping Study. *Sensors* 23, 19 (2023). <https://doi.org/10.3390/s23198056>
- [36] Joosep Parts, Janika Leoste, Kalle Tammemäe, and Slavko Rakic. 2025. A Systematic Scoping Review of Privacy Challenges and Privacy Enhancing Technologies in Teleoperated Robotics. *IEEE Access* 13 (2025), 216724–216747. <https://doi.org/10.1109/ACCESS.2025.3645240>
- [37] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. In *Annual Meeting of the Association for Computational Linguistics*. <https://api.semanticscholar.org/CorpusID:254854519>
- [38] Maria Pinto, Matthijs Biondina, and Tony Belpaeme. 2025. Designing Social Robots with LLMs for Engaging Human Interaction. *Applied Sciences* 15 (2025), 6377. <https://doi.org/10.3390/app15116377>
- [39] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084 [cs.CL] <https://arxiv.org/abs/1908.10084>
- [40] Research and Markets. 2025. *Social Robots Market Report 2025*. Technical Report. Research and Markets. <https://www.researchandmarkets.com/reports/5930943/social-robots-market-report> Accessed: 2025.
- [41] Gregory Schwartzman. 2024. Exfiltration of Personal Information from ChatGPT via Prompt Injection. arXiv:2406.00199 [cs.CR] <https://arxiv.org/abs/2406.00199>
- [42] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duveaud, Amanda Askill, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, et al. 2025. Towards Understanding Sycophancy in Language Models. arXiv:2310.13548 [cs.CL] <https://arxiv.org/abs/2310.13548>
- [43] Zhonghao Shi, Ellen Landrum, Amy O'Connell, Mina Kian, Leticia Pinto-Alva, Kaleen Shrestha, Xiaoyuan Zhu, and Maja J. Mataric. 2024. How Can Large Language Models Enable Better Socially Assistive Human-Robot Interaction: A Brief Survey. arXiv:2404.00938 [cs.HC] <https://arxiv.org/abs/2404.00938>
- [44] Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, et al. 2026. Emotion Concepts and Their Function in a Large Language Model. arXiv:2604.07729 [cs.AI] <https://arxiv.org/abs/2604.07729>
- [45] Julia Stapels, Angelika Augustine, Niels Diekmann, and Friederike Eyssel. 2023. Never Trust Anything That Can Think for Itself, if You Can't Control Its Privacy Settings: The Influence of a Robot's Privacy Settings on Users' Attitudes and Willingness to Self-disclose. *International Journal of Social Robotics* 15 (09 2023), 1–19. <https://doi.org/10.1007/s12369-023-01043-8>
- [46] Erich Studerus, Vivienne Jia Zhong, and Stephan Vonschallen. 2026. A Framework for Low-Latency, LLM-Driven Multimodal Interaction on the Pepper Robot. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI '26)*. ACM, 1298–1302. <https://doi.org/10.1145/3757279.3788808>
- [47] Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long T. Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. 2025. In Prospect and Retrospect: Reflective Memory Management for Long-term Personalized Dialogue Agents. arXiv:2503.08026 [cs.CL] <https://arxiv.org/abs/2503.08026>
- [48] Wenxiao Zhang, Xiangrui Kong, Conan Dewitt, Thomas Braunl, and Jin B. Hong. 2024. A Study on Prompt Injection Attack Against LLM-Integrated Mobile Robotic Systems. arXiv:2408.03515 [cs.RO] <https://arxiv.org/abs/2408.03515>
- [49] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. It's a Fair Game, or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 1–26. <https://doi.org/10.1145/3613904.3642385>
- [50] Wenzheng Zhao, Kruthika Gangaraju, and Fengpei Yuan. 2025. Multimodal perception-driven decision-making for human-robot interaction: a survey. *Frontiers in Robotics and AI* Volume 12 - 2025 (2025). <https://doi.org/10.3389/frobt.2025.1604472>