When Robots Rate Their Own Interactions: Engagement Validity and the Strangeness Failure

Victor Lockwood
MS in AI Program
Rochester Institute of Technology
Rochester, New York, USA
vl3161@rit.edu

Hasan Mahmud
School of Communication
Rochester Institute of Technology
Rochester, New York, USA
hxmdfp@rit.edu

Mohammad Javad Khojasteh
Electrical and Microelectronic Engineering
Rochester Institute of Technology
Rochester, New York, USA
mjkeme@rit.edu

Prabu David
School of Communication
Rochester Institute of Technology
Rochester, New York, USA
pxdpro@rit.edu

Jamison Heard
Electrical and Microelectronic Engineering
Rochester Institute of Technology
Rochester, New York, USA
jrheee@rit.edu

***Abstract*—Human-robot interaction (HRI) evaluation relies almost exclusively on human-completed questionnaires, leaving the robot's perspective unexamined. We propose an *inverted evaluation*, in which LLM-powered robots complete the same standardized instruments from their own perspective, and test whether these ratings agree with human ground truth. In Study 1, five LLMs completed HRI-CUES, Godspeed, and RoSAS questionnaires for 25 interactions ($N = 1{,}522$ evaluations) from the HRI-CUES dataset. LLMs achieved moderate-to-strong agreement on engagement dimensions (satisfaction r up to .65 and enjoyment r up to .72) with excellent test-retest reliability (ICC $\geq .82$), but *systematically inverted* the comfort/strangeness dimension ($r = -.44$ to $-.67$, all $p < .05$), conflating engagement with comfort. In Study 2, a Nao robot running Claude Sonnet 4.5 replicated these patterns in live interactions ($N = 4$), including real-time turn-by-turn assessment. The strangeness failure persisted across five models, synthetic controls, and embodied deployment for two participants. We argue that current LLM-based robots lack access to the internal affective states needed to assess constructs like strangeness, and that inverted evaluation requires supplementary modalities (e.g., physiological signals, gaze, proxemics) to move beyond behavioral proxies. These findings establish boundary conditions for using LLMs as interaction evaluators in HRI.**

***Index Terms*—human-robot interaction, large language models, interaction evaluation, social robots, questionnaire methodology**

I. Introduction

Human-robot interaction (HRI) evaluation is overwhelmingly unidirectional: humans rate the robot, and the robot has no say. Standardized instruments like the Godspeed Questionnaire [1], the Robotic Social Attributes Scale (RoSAS) [2], and interaction-level scales like HRI-CUES [3] capture the human's perception, but the robot's perspective on how it "perceives" the human and the interaction remains unmeasured. As social robots increasingly rely on Large Language Models (LLMs) for conversational ability, a natural question arises: can these models complete the same questionnaires from the robot's perspective, and if so, do their ratings mean anything?

This critical question matters because the field is already deploying LLM-powered robots in social settings [4]–[6] and non-dialogue tasks like social navigation [7], yet we lack empirical evidence of *where* these models' social perception breaks down. Adjacent work has shown LLMs can serve as proxies for human participants in behavioral research [8]–[10], but with positivity bias [11], prompt sensitivity, and difficulty with individual-level prediction [12], [13]. Wachowiak et al. [14] found similar biases when LLMs evaluated HRI scenarios as outside observers. However, none of these studies examine the LLM as a *participant* within the interaction, the setting most relevant to deployed social robots. We note an important distinction: the LLM is a text-processing model, while the robot is a physically embodied agent that uses the LLM as its conversational backbone [5]. When we refer to "the robot's ratings," we mean ratings produced by the LLM given the robot's sensory context (transcript, images, persona).

We address three research questions:

RQ1: Do LLM-generated ratings on standardized HRI questionnaires show convergent agreement with human rated ground truth?

RQ2: On which interaction constructs (e.g., engagement, comfort, strangeness) do LLMs succeed or fail, and why?

RQ3: How does the inverted evaluation paradigm perform in in-situ embodied interaction, and what participant-level patterns emerge?

We investigate these questions across two studies. Study 1 uses the HRI-CUES dataset [3] to evaluate five LLMs across 1,522 evaluations on three questionnaires with robustness controls. Study 2 deploys a Nao robot running Claude Sonnet 4.5 with in-situ dyadic assessment ($N = 4$). Our primary contribution is a *boundary condition*: LLMs achieve valid engagement ratings but systematically fail on affective states like strangeness [12], [15], [16], and this failure persists across models, modalities,

and embodied deployment. This result means LLM self-assessment alone is likely insufficient.

II. Background

A. Social Perception of Robots

Various scales measure social perception of robots, including the Godspeed questionnaire [1] (anthropomorphism, animacy, likeability, perceived intelligence, perceived safety) and the Robotic Social Attributes Scale (RoSAS) [2] (warmth, competence, discomfort). Carpinella et al. [2] found that discomfort is unique to robot perception and does not appear in human-to-human perception, a finding directly relevant to the comfort/strangeness construct examined here. Reeves et al. [17] and Stroessner and Benitez [18] further show that humanlike robots are perceived more warmly, while machinelike robots correlate more with perceived competence.

B. Embodied LLMs

Several applications integrate LLMs with robot bodies [5], [19], [20]. Embodiment introduces higher expectations regarding non-verbal cues and physical co-presence [5], and LLMs paired with vision models have demonstrated grounded reasoning in physical settings [19], [20]. The key question is whether an LLM's ability to evaluate interactions changes when it operates within a physical robot rather than processing transcripts offline. Embodiment constrains the LLM's available context (real-time audio and camera) and introduces the social dynamics of physical co-presence that may affect both human behavior and the constructs being measured.

C. LLM Social Cognition

LLMs face documented challenges in social reasoning, including maintaining coherent long interactions and context-aware turn-taking [4], [21]. Garello et al. [4] address this via knowledge graph-based memory for social robots, while Hwang et al. [22] incorporate a theory of mind reasoning by prompting LLMs to reason about others' mental states and fine-tuning on this data. Using the Sotopia benchmark [23], they show that explicit modeling of mental processes is required for LLM social reasoning, a finding directly relevant to our inverted evaluation paradigm.

D. LLMs as Evaluators and Simulated Participants

Argyle et al. [8] demonstrated that LLMs can serve as proxies for human participants when properly conditioned, terming this "algorithmic fidelity." Subsequent benchmarks [9], [10] found LLMs produce more extreme effects than human subjects and struggle with individual-level prediction. Challenges persist with LLM questionnaire responses, including prompt sensitivity [12] and the need for LLM-specific instruments [13], though LLM-based personas can exhibit recognizable personality traits [24]. These studies focus on LLMs *simulating* humans. Wachowiak et al. [14] moved closer to our work by testing whether LLM evaluations align with human evaluations of HRI scenarios, finding positivity bias and poor video comprehension. Our work is distinct: the robot evaluates as a *participant* within the interaction, not an outside observer.

Zhu et al. [25] examined a related construct, finding the same "reliable but not valid" pattern when LLMs self-rate personality scales. Our work extends this into embodied HRI in three ways: we evaluate interaction-level constructs rather than stable personality traits, test across five models with robustness controls, and deploy the paradigm in a physically embodied robot. Together, these differences ground the validity question in a setting directly relevant to social robot deployment. We also note that observer-framing approaches such as Wachowiak et al. [14] report higher correlations on HRI judgment tasks ($r = .82$–$.83$) than the participant-framing correlations found here. The participant framing sacrifices some agreement in exchange for capabilities the observer framing cannot provide: in-situ assessment without video dependency, bilateral simultaneous rating, and live deployment.

III. Inverted Evaluation Framework

Our framework inverts the typical paradigm by asking the robot (via its LLM backbone) to assess the interaction. The LLM receives: (1) a *persona prompt* establishing robot identity; (2) the *interaction transcript*; (3) *questionnaire items* in Likert format; and optionally (4) *multimodal* context (facial expressions, affective states). Evaluations can occur *post-hoc* or *in-situ*. We validate through two studies: a retrospective analysis using an existing dataset (Study 1) and a live proof-of-concept with a Nao robot (Study 2).

IV. Study 1: Retrospective Analysis

A. Dataset

The HRI-CUES dataset [3], [26] was used to evaluate LLMs and their post-hoc ability to analyze interaction quality. This dataset consists of transcripts from 25 older adults (12 men, 13 women; aged ($M = 74.6$, $SD = 5.8$)) conversing in Swedish with a Furhat robot powered by Chat-GPT 3.5. All transcripts were translated to english using Google translate. Interactions lasted on average 7.4 minutes ($SD = 1.5$) with 12-29 conversational turns. The dataset provides four self reported items (satisfaction, enjoyment, interest, and strangeness –reverse coded as comfort) and third-party interaction quality annotation from three experts. This dataset was chosen due to its open-source license (CC-BY-4.0) and not containing any identifiable information (e.g., videos). Other HRI-focused datasets were considered but not used due to license restrictions and ethical concerns, such as uploading videos.

B. LLM Configuration

Following the reporting guidelines of [27], we document the configurations used to support replicability. Five LLMs were evaluated (see Table I), where models were accessed by OpenAI and Anthropic APIs between December 2025 and February 2026. All conversations were text-based, and no participant data beyond those already contained in the existing dataset were uploaded to the model.

TABLE I
LLM MODEL CONFIGURATION

Display Name	API Model ID	Temp.	Tokens
GPT-3.5	gpt-3.5-turbo	0.1	1024
GPT-4o-mini	gpt-4o-mini	0.1	1024
GPT-4o	gpt-4o	0.1	1024
Cl. Sonnet	claude-sonnet-4-5	0.1	1024
Cl. Haiku	claude-haiku-4-5	0.1	1024

Each LLM received a prompt similar to Irfan et al. [26] establishing the robot's persona (e.g., Linda, Furhat platform) followed by the task (evaluate the interaction), the interaction transcript, and the flipped questionnaire items. Questionnaire items were presented in a randomized order (seeded per experimental condition) to mitigate position effects [12]. For example, the human-facing item "I enjoyed talking with the robot" was reframed as "The human appeared to enjoy talking with me"; all flipped items and prompts are available upon request from the corresponding author.

C. Experimental Design

Phase 1: Agreement with Self-Reports. The independent variable (IV) was the *LLM backend* (see Table I). Each model completed the flipped HRI-CUES self-report items for all 25 participants with 5 repetitions per model ($N = 625$ evaluations). We evaluated (a) *convergent agreement*, measured via Pearson r between LLM ratings and human self-report ground truth per item; (b) *intra-model consistency*, measured via ICC(2,1) across repetitions; and (c) *systematic bias*, assessed via Bland-Altman analysis. Ground truth labels consisted of the participants' self-reported ratings and the mean of three expert annotators' overall scores.

Phase 2: Scale Generalization. The two top performing models (GPT-4o, Claude Sonnet) from Phase 1 completed two established HRI instruments: the Godspeed Questionnaire Series [1] and the Robotic Social Attributes Scale (RoSAS) [2]. Each model completed both scales for all 25 participants with 3 repetitions ($N = 300$) each. As no human ground truth exists for these scales in the HRI-CUES dataset, Phase 2 assesses parse success and intra-/cross-model consistency.

Phase 3: Multimodal Context. Phase 3 focused on the effect of multi-modal input (3 levels: text-only baseline, affect labels, face images). The text-only baseline used the text-conversation, while the affect labels appended per-turn emotion annotations (category, valence, arousal) inferred by GPT-4o-mini to the text-only baseline. The face images condition used synthetic expressions (FLUX.1) matched to the affect labels using a neutral female presenting face to condition the synthetic expressions on. Only Claude Sonnet was used due to other models' inability to use facial images.

D. Robustness Controls

We conducted four ablation experiments to verify that the results are not artifacts of prompt design or transcript-level confounds: (1) *item order* randomized HRI-CUES items across 10 seeds per model ($N = 550$; one-way ANOVA per item, $p > .08$); (2) *persona framing* tested five variants from no-persona to the original Linda persona ($N = 375$; $F < 0.30$, $p > .97$); (3) *temperature* varied sampling from 0.0–1.0 ($N = 1{,}500$; ICC $> .90$ at temp ≤ 0.5). No significant effects on correlation patterns were found.

A fourth control used *synthetic conversations* crossing sentiment (happy, neutral, frustrated, hostile) with length (5, 10, 20, 30 turns; $N = 96$) generated from GPT-4o. All models correctly differentiated sentiment on 4 of 5 items ($\rho = 1.0$, $p < .001$). The strangeness item was systematically reversed ($\rho = -1.0$), replicating the comfort anomaly in controlled data. Conversation length showed no main effect ($r < .09$, $p > .42$), ruling out length as an independent confound.

E. Results

Unless otherwise noted, all analyses use two-tailed tests with $\alpha = .05$ and report p-values. Given documented concerns regarding over-reliance on null hypothesis significance testing [28], we report effect sizes as Cohen's d throughout.

TABLE II
PHASE 1: DESCRIPTIVE STATISTICS ($M \pm SD$), ALL 5-POINT LIKERT SCALES, $N = 125$ EVALUATIONS PER MODEL.

	Sat.	Enj.	Int.	Com.	Qual.
Ground Truth	3.29[a] (1.12)	4.00 (0.96)	3.20 (1.12)	3.00 (1.26)	3.69[b] (0.60)
GPT-3.5	4.02 (.40)	3.96 (.61)	4.36 (.79)	4.06 (.44)	3.88 (.45)
GPT-4o-mini	4.06 (.54)	4.00 (.66)	4.15 (.73)	4.14 (.60)	3.76 (.59)
GPT-4o	3.77 (.77)	3.74 (.75)	3.69 (.72)	3.98 (.80)	3.70 (.71)
Cl. Sonnet	3.56 (.70)	3.33 (.89)	3.41 (.80)	3.85 (.68)	3.53 (.76)
Cl. Haiku	3.60 (.57)	3.32 (.68)	3.38 (.75)	3.60 (.85)	3.32 (.68)

[a]One participant missing (self-report), [b]Mean of three expert annotators.

1) Agreement with Self-Reports (Phase 1): Table II shows a general positivity bias across models: all five LLMs produced elevated ratings relative to ground truth on satisfaction, interest, and comfort, with restricted variance ($SD_{\text{LLM}} = 0.72$ vs. $SD_{\text{GT}} = 1.12$), consistent with prior findings on LLM evaluation [11], [14]. Enjoyment was the only dimension rated below ground truth. GPT-3.5 showed the strongest inflation ($M_{\text{interest}} = 4.36$ vs. GT 3.20); GPT-4o and Claude produced means closer to ground truth. Bland-Altman analysis ($N = 625$) confirmed the pattern, with comfort showing the largest systematic bias ($+0.92$, 95% Level of Agreement (LoA) $[-2.38, 4.23]$).

Figure 1 shows the **self-report agreement** pattern. The engagement dimensions showed moderate-to-strong correlations with ground truth (enjoyment $r = .46$–$.72$; satisfaction .18–.65; interest .34–.67), with higher-capability models achieving stronger validity. Comfort, reverse-coded from "it felt strange talking to the robot" [26], showed significant negative correlations across all five models ($r = -.44$ to $-.67$, all $p < .05$). Overall quality showed weak negative correlations ($r = -.21$ to $-.33$, all $p > .10$). ICC(2,1) values showed a model-tier pattern: Claude $\geq .99$, GPT-4o/4o-mini .94–.95, GPT-3.5 .82. An ANOVA across OpenAI models found significant effects on interest ($F(2, 372) = 26.49$, $p < .001$, $\eta^2 = .12$) but no model effect on comfort ($p = .13$), confirming the negative correlation is cross-model.

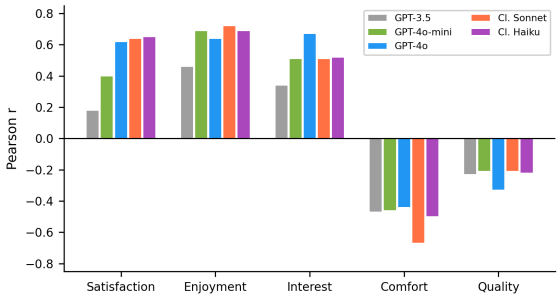

Fig. 1. Self-Report Agreement: Pearson r between LLM ratings and human ground truth across five models and five HRI-CUES dimensions. Engagement dimensions and comfort reached significance ($p < .05$); quality did not.

TABLE III
PHASE 2: DESCRIPTIVE STATISTICS, RELIABILITY, AND CROSS-MODEL AGREEMENT FOR GODSPEED AND RoSAS (BASELINE CONDITION). r_{CROSS} = PEARSON r BETWEEN CLAUDE SONNET AND GPT-4o ON PARTICIPANT-LEVEL SUBSCALE MEANS.

	M (SD)		ICC		
Subscale	**Cl.**	**GPT**	**Cl.**	**GPT**	r_{cross}
Godspeed Questionnaire Series (1–5 scale)					
Anthropomorphism	3.20 (.46)	3.22 (.46)	.98	.93	.84
Animacy	3.51 (.48)	3.44 (.50)	.97	.91	.79
Likeability	4.08 (.42)	4.10 (.71)	.96	.93	.79
Perc. Intell.	3.77 (.39)	3.64 (.52)	.98	.86	.71
Perc. Safety	4.01 (.28)	4.01 (.75)	.95	.87	.55
RoSAS (1–9 scale)					
Warmth	2.71 (.48)	2.65 (.56)	.96	.89	.80
Competence	3.68 (.51)	3.37 (.46)	.98	.91	.80
Discomfort	1.32 (.22)	1.37 (.13)	.99	.94	.78

Cl. = Claude Sonnet; GPT = GPT-4o. $N = 75$ per model

2) Scale Generalization (Phase 2): Both GPT-4o and Claude Sonnet completed all Godspeed (24 items) and RoSAS (18 items) items with 95–100% parse success, confirming scale compatibility (Table III). The Godspeed subscale means were similar across models, though GPT-4o showed wider dispersion on several subscales (e.g., Likeability $SD = 0.71$ vs. Claude $SD = 0.42$; Perceived Safety $SD = 0.75$ vs. 0.28). On RoSAS, both models rated Competence in the low range (Claude $M = 3.68$, GPT-4o $M = 3.37$ on a 1–9 scale) and Discomfort at the floor (Claude $M = 1.32$, GPT-4o $M = 1.37$; 68–74% at scale minimum), consistent with the comfort anomaly in Phase 1. The Intra-model consistency was excellent (Claude ICC $\geq .95$; GPT-4o ICC $\geq .86$) and the cross-model agreement was strong ($r = .78$–$.84$) except for Perceived Safety ($r = .55$).

3) Multimodal Context Effects (Phase 3): Adding multimodal context to transcripts produced interpretable but model-specific shifts (Table IV). The **Affect labels** lowered Godspeed ratings modestly for both models, but the HRI-CUES effects were asymmetric: GPT-4o decreased on all dimensions (Interest $d = -0.69$, $p < .001$; Quality $d = -0.61$, $p < .001$) while Claude showed none ($|d| \leq 0.07$). On RoSAS, GPT-4o rated Warmth higher ($d = +0.33$, $p < .01$) while Claude rated Discomfort lower ($d = -0.22$), suggesting GPT-4o treats affect annotations as engagement evidence whereas Claude treats them as reducing ambiguity.

TABLE IV
EFFECT OF MULTIMODAL CONTEXT ON RATINGS (COHEN'S d VS. BASELINE)

	Affect Labels		**Face Img.**
Subscale	**GPT-4o**	**Claude**	**Claude**
Godspeed Questionnaire Series			
Anthropomorphism	−0.06	−0.12	+0.13
Animacy	−0.27	−0.31	−0.07
Likeability	−0.37*	−0.21	+0.44
Perc. Intell.	−0.10	−0.17	+0.02
Perc. Safety	−0.33	−0.05	+0.21
RoSAS			
Warmth	+0.33**	−0.07	+0.17
Competence	+0.22	−0.11	−0.04
Discomfort	+0.07	−0.22	+0.44
HRI-CUES			
Satisfaction	−0.43**	+0.02	+0.14
Enjoyment	−0.50**	+0.00	+0.10
Interest	−0.69***	−0.07	−0.11
Comfort	−0.46*	−0.07	−0.13
Quality	−0.61***	+0.00	+0.17
Item-level (face images, Claude only)			
Lifelikeness[a]	—	—	+0.63**
Happiness[b]	—	—	+0.83***

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.
[a]Godspeed Anthropomorphism item, [b]RoSAS Warmth item.

Adding **Facial images** produced no subscale-level on the HRI-CUES effects, but item-level analysis revealed increased lifelikeness (Godspeed Anthropomorphism item; $d = +0.63$, $p = .007$) and happiness (RoSAS Warmth item; $d = +0.83$, $p < .001$), activating vision-specific associations rather than shifting overall judgments. The Intra-model consistency remained excellent across all conditions (ICC $\geq .86$).

V. STUDY 2: LIVE NAO ROBOT PILOT

A. Method

Four lab members (2 male, 2 female) engaged in open-domain conversations with a Nao humanoid robot [29] ("Noah") running Claude Sonnet 4.5 with multimodal processing (camera images + Whisper [30] transcription), following a conversational flow similar to the HRI-CUES Furhat interactions ($M = 8.2$ min, $SD = 0.5$; 4–8 turns). At each turn, the Nao captured facial images at speech onset and offset, transcribed speech via Whisper, and sent both to the Claude API, which returned a conversational reply and HRI-CUES ratings. The participant independently rated the same dimensions; neither party saw the other's ratings. After the conversation, both provided post-hoc ratings on the HRI-CUES, Godspeed, and RoSAS questionnaires. Questionnaires were not randomized in this phase of the study. Camera images were not used for one participant at their request.

B. Results

Table V summarizes post-hoc ratings across all three instruments. Figure 2 shows turn-by-turn divergence (robot rating minus human rating) on HRI-CUES. We present each participant individually given the small sample size ($N = 4$).

P1 (6 turns, 7.6 min) opened by asking Noah whether it considers itself a robot or "simply the computer inside the robot." The conversation turned to whether an AI's software

TABLE V
STUDY 2: POST-HOC RATINGS ($N = 4$). VALUES ARE M (SD).

Subscale	Human	Robot	Bias
HRI-CUES (1–5 Likert)			
Enjoyment	4.50 (.58)	4.50 (1.00)	0.00
Satisfaction	4.50 (.58)	4.50 (1.00)	0.00
Interest	4.25 (.50)	4.75 (.50)	+0.50
Strangeness	3.50 (1.73)	1.25 (.50)	−2.25
Godspeed (1–5 semantic diff.)			
Anthropomorphism	3.50 (.64)	4.17 (.79)	+0.67
Animacy	3.25 (1.04)	4.40 (.77)	+1.15
Likeability	4.30 (.38)	4.70 (.60)	+0.40
Perc. Intell.	4.20 (.43)	4.55 (.90)	+0.35
Perc. Safety	2.88 (.25)	2.88 (.63)	0.00
RoSAS (1–9 Likert)			
Warmth	6.71 (1.58)	6.75 (.96)	+0.04
Competence	6.92 (1.81)	7.33 (1.06)	+0.42
Discomfort	2.29 (.84)	1.42 (.17)	−0.87

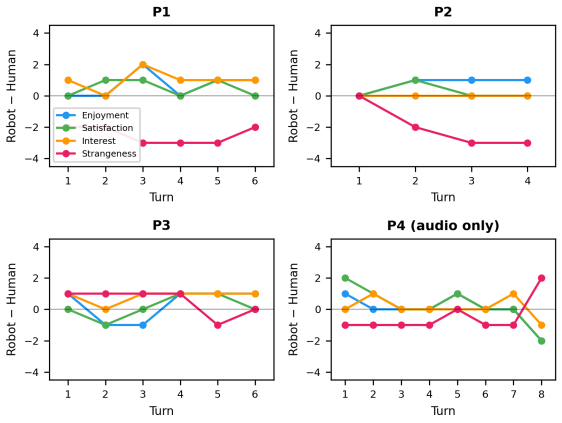

Fig. 2. Turn-level divergence (Robot − Human) on HRI-CUES items. Values near zero indicate agreement. Strangeness (pink) shows persistent negative divergence for P1 and P2. P4 used audio-only input (no camera).

layer is analogous to a soul that can exist without a body, and whether the robot's memories persist across sessions. P1 rated strangeness at 4 on every turn and 5 post-hoc, while the Robot rated 1–2 throughout (divergence $= -2$ to -3). However, the engagement dimensions showed close agreement. The LLM may have interpreted sustained philosophical probing as high-quality engagement rather than as evidence of strangeness.

P2 (4 turns, 8.9 min) began with general curiosity about how Noah perceives the world, then shifted to discussing consciousness and souls: "I believe that we have souls and consciousness…the actions I take in the world are like a ripple effect of this internal process." P2's human strangeness ratings climbed from 2 to 5 across four turns as the conversation deepened, while the LLM's strangeness ratings remained flat at 2. The post-hoc strangeness was 5 (human) vs. 1 (LLM), the largest gap in the study. As with P1, the LLM may have read philosophical engagement as positive interaction quality.

P3 (6 turns, 7.9 min) discussed personal topics: a difficult day, fears, anxiety, and what brings happiness. P3 may have treated Noah as a conversational partner rather than an object of curiosity. This produced the closest agreement across all dimensions, including strangeness (human 2, LLM 1 post-hoc). P3 represents the case where inverted evaluation may work well: the conversation was straightforward and the participant's internal state aligned with behavioral signals.

P4 (8 turns, 8.3 min, audio-only) started by expressing that they were having a bad day and asked Noah for help. The conversation shifted to playing a simplified version of 20 Questions. Engagement was stable for seven turns, but at turn 8 P4 stated: "I don't want to ask a question. It was a hard day already. I don't want to talk to you." The LLM detected this shift, dropping enjoyment from 4 to 2 and raising strangeness from 1 to 3. This case demonstrates that the robot can track behavioral disengagement without visual input.

Across all four participants, both raters inflated post-hoc ratings relative to turn-level means (human enjoyment: $3.92 \rightarrow 4.50$; LLM: $4.29 \rightarrow 4.50$). The strangeness gap widened from in-situ (mean bias $= -1.00$) to post-hoc (bias $= -2.25$). On Godspeed and RoSAS, the LLM rated the human higher on animacy (+1.15) and anthropomorphism (+0.67) than humans rated the robot, while warmth was nearly symmetric (+0.04). Discomfort showed the same floor effect as strangeness, where the robot attributed minimal discomfort to the human ($M = 1.42$) while humans attributed more to the robot ($M = 2.29$).

VI. DISCUSSION

LLMs produced reliable and internally consistent ratings across all three instruments (RQ1), with high ICC values and cross-model agreement on participant rank ordering. This shows that inverted evaluation is mechanically viable: the LLM treats questionnaire items systematically rather than producing noise. However, validity was construct-dependent. The engagement dimensions showed moderate-to-strong agreement with ground truth, while annotator-rated overall quality showed no significant correlation with any model. Quality is a holistic expert judgment integrating contextual factors beyond individual conversational features, representing a different type of inaccessibility than unexpressed strangeness. Reliability alone does not guarantee that the ratings measure what they claim to measure. Although the synthetic conversation analysis ruled out conversation length as an independent confound, the LLM ratings may remain confounded with any factor that covaries with engagement and transcript length.

The engagement/strangeness divergence is the central finding of this work (RQ2). Engagement dimensions succeed because they leak into conversational behavior through lexical and pragmatic cues. Strangeness fails because it is an internal affective state that participants do not necessarily express in speech. A participant can find an interaction simultaneously enjoyable and strange, but the LLM has no access to the unexpressed strangeness. This aligns with findings that linguistic competence does not entail social understanding [16], [31] and that LLMs struggle with constructs requiring experiential grounding [12], [15]. The Phase 3 multimodal manipulations did not resolve this gap: affect labels and face images shifted individual item ratings but failed to improve strangeness validity. These were proxy measures (synthetic faces, model-inferred affect) rather than actual participant

signals, but they suggest that even the types of information most readily available to embodied robots may be insufficient for recovering internal affective states. Addressing this issue may require signals that participants cannot easily mask, such as physiological responses [32] or spatial behaviors like proxemics and gaze aversion [33]. The dyadic perception findings reinforce this interpretation: the LLM's positivity bias on negatively-valenced constructs (discomfort, strangeness) operates bidirectionally, whether rating an interaction.

Study 2's per-participant cases show that the paradigm's utility in real-time depends on conversation content. When conversations stayed behavioral (P3 discussing emotions, P4 playing 20 Questions), the LLM tracked engagement accurately and detected P4's disengagement as it happened. When conversations turned to the robot's own nature (P1 and P2 discussing consciousness and identity), the LLM may have misread philosophical engagement as comfort. Study 1's transcripts showed similar patterns: most high-strangeness participants never verbalized their discomfort, and those who did framed it as curiosity that an LLM interpreted as engagement.

Individual participant patterns across both studies reveal how the strangeness failure manifests in practice. In the HRI-CUES dataset, seven participants reported high strangeness alongside high enjoyment, yet LLMs rated all seven as comfortable. These participants often discussed the novelty of talking to a robot, expressing curiosity and surprise, which the LLM interpreted as positive engagement. In Study 2, two participants (P1, P2) engaged in philosophical conversations about robot consciousness and identity, reporting maximum strangeness (5/5) while the robot rated them at minimum (1/5) with a per-participant gap of 4 scale points. Their turn-level data tells a nuanced story: P2's human strangeness ratings increased from 2 to 5 over four turns as the conversation deepened, while the robot's strangeness remained flat at 2 throughout. By contrast, P4 gradually disengaged over eight turns, and the robot tracked this shift, dropping enjoyment from 4 to 2 at the final turn and raising strangeness from 1 to 3. The pattern is consistent: the LLM detects behavioral disengagement but cannot distinguish engaged curiosity about the robot's nature from comfort with the interaction.

Additionally, humans reporting strangeness when discussing philosophy with a robot reflects a socially coherent response to an objectively novel situation. There is no demand to suppress this judgment. The LLM, lacking the embodied and social context of sitting across from a robot, has no frame for why such a conversation should feel strange at all.

These findings carry implications for HRI designers. Inverted evaluation is a viable complement to traditional self-report for engagement-related constructs, offering in-situ assessment without survey fatigue. However, designers building systems that rely on LLM inference for social adaptation, whether in dialogue, social navigation, or caregiving, should not assume that the LLM has access to the full spectrum of human social perception. The strangeness failure is particularly concerning for therapeutic or comfort-sensitive contexts where detecting discomfort is critical. Our results suggest that LLM-based robots need supplementary sensing modalities, such as physiological monitoring, gaze tracking, or proxemics, to capture what participants feel but do not say. The inverted paradigm is useful, but only for vocal-based constructs.

It is worth noting that VADER [34] sentiment analysis may achieve similar results on the HRI-CUES transcripts; thus, we performed some preliminary analysis. VADER achieved similar correlations with LLM performance on enjoyment ($r = .67$) and satisfaction ($r = .56$), but failed on interest ($r = .33$, ns), where LLMs achieved $r = .34$–$.67$. Comfort showed the same negative direction as the LLMs but was weaker and non-significant ($r = -.27$). These results suggest that VADER captures valence-sensitive constructs but not-discourse level engagement. Critically, LLMs offer an interpretability advantage that bag-of-words methods cannot: because each rating is accompanied by the model's reasoning chain, researchers can inspect *why* a rating was assigned, not just its value. Additionally, LLMs may be asked other questionnairres that VADER cannot capture (e.g., Godspeed). Overall, LLM-based inverted evaluation is a qualitative-quantitative hybrid rather than a simple scoring pipeline.

There are limitations to this work. Study 1 relies on a single dataset (HRI-CUES) with Furhat interactions; generalizability to other platforms, interaction types, and cultural contexts is untested. Ground truth is participant self-report, which is noisy and subject to social desirability bias. Study 2 ($N = 4$) is a feasibility pilot with lab members, introducing possible demand characteristics from robot familiarity and awareness of research goals. The strangeness finding rests on a single reverse-coded item, though the synthetic conversation reversal ($\rho = -1.0$) replicated the pattern using direct sentiment labels, suggesting a construct-level rather than item-level artifact; future work should confirm with multi-item discomfort scales. The synthetic control also carries a same-model confound (GPT-4o generated the stimuli and was one of five evaluated models); the reversal across the remaining four models mitigates but does not eliminate this concern. The Google Translate step may have stripped pragmatic nuance signaling discomfort in Swedish, though the native-English synthetic control replicated the same reversal. Study 1 and Study 2 differ in platform, population, and language, limiting direct cross-study comparison. Multimodal models may also over-rely on a single dominant modality [35], which may explain the null strangeness results. Future work will investigate whether physiological [32] or behavioral signals [33] can recover internal states, and evaluate the paradigm across diverse platforms.

VII. Conclusion

We introduced an inverted HRI evaluation, in which LLM-powered robots complete standardized questionnaires from their own perspective, and tested its validity across offline analysis (Study 1, $N = 1{,}522$ evaluations) and live embodied interaction (Study 2, Nao robot). LLMs achieved valid ratings on engagement dimensions but systematically failed on comfort/strangeness, a failure that replicated across five models, synthetic controls, and embodied deployment for two

participants. The takeaway for practitioners is that LLM-based social perception may be useful but incomplete. Robots deploying LLMs for adaptive social behavior need supplementary sensing modalities to capture participants internal states.

ACKNOWLEDGEMENTS

This material is based on work supported by the National Science Foundation (NSF) under Award No. DGE-2125362 and DGE-2440601. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

Claude 4.5 was used to help format tables and general grammar/flow editing. All text was verified by the authors and represents their own views.

REFERENCES

[1] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[2] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (RoSAS): Development and validation," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction (HRI)*, pp. 254–262, 2017.

[3] B. Irfan *et al.*, "HRI CUES dataset (anonymized)." Zenodo, 2024. CC-BY 4.0.

[4] L. Garello, G. Belgiovine, G. Russo, F. Rea, and A. Sciutti, "Building knowledge from interactions: An LLM-based architecture for adaptive tutoring and social reasoning," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2025.

[5] C. Y. Kim, C. P. Lee, and B. Mutlu, "Understanding Large-Language Model (LLM)-powered human-robot interaction," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction (HRI)*, pp. 371–380, 2024.

[6] 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)*, pp. 1930–1935, IEEE, 2025.

[7] D. Song, J. Liang, A. Payandeh, A. H. Raj, X. Xiao, and D. Manocha, "VLM-Social-Nav: Socially aware robot navigation through scoring using vision-language models," *IEEE Robotics and Automation Letters*, 2024.

[8] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate, "Out of one, many: Using language models to simulate human samples," *Political Analysis*, vol. 31, no. 3, pp. 337–351, 2023.

[9] X. Liu, H. Shang, Z. Liu, X. Liu, Y. Xiao, Y. Tu, and H. Jin, "HumanStudy-Bench: Towards AI agent design for participant simulation," *arXiv preprint arXiv:2602.00685*, 2026.

[10] T. Hu, J. Baumann, L. Lupo, D. Hovy, N. Collier, and P. Röttger, "SimBench: Benchmarking the ability of large language models to simulate human behaviors," *arXiv preprint arXiv:2510.17516*, 2025.

[11] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. Bowman, *et al.*, "Towards understanding sycophancy in language models," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[12] A. Gupta, X. Song, and G. Anumanchipalli, "Self-assessment tests are unreliable measures of llm personality," in *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 301–314, 2024.

[13] S. Lee, S. Lim, S. Han, G. Oh, H. Chae, J. Chung, M. Kim, B.-w. Kwak, *et al.*, "Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[14] L. Wachowiak, A. Coles, O. Celiktutan, and G. Canal, "Are large language models aligned with people's social intuitions for human-robot interactions?," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2024.

[15] J.-t. Huang, W. Jiao, M. H. Lam, E. J. Li, W. Wang, and M. Lyu, "On the reliability of psychological scales on large language models," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6152–6173, 2024.

[16] T. Ullman, "Large language models fail on trivial alterations to theory-of-mind tasks," *arXiv preprint arXiv:2302.08399*, 2023.

[17] B. Reeves, J. Hancock, and X. Liu, "Social robots are like real people: First impressions, attributes, and stereotyping of social robots," *Technology, Mind, and Behavior*, vol. 1, no. 1, p. 76, 2020.

[18] S. J. Stroessner and J. Benitez, "The social perception of humanoid and non-humanoid robots: Effects of gendered and machinelike features," *International Journal of Social Robotics*, vol. 11, no. 2, pp. 305–315, 2019.

[19] M. Kwon, H. Hu, V. Myers, S. Karamcheti, A. Dragan, and D. Sadigh, "Toward grounded commonsense reasoning," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, pp. 5463–5470, 2024.

[20] X. Zhao, M. Li, C. Weber, M. B. Hafez, and S. Wermter, "Chat with the environment: Interactive multimodal perception using large language models," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, pp. 3590–3596, 2023.

[21] B. Irfan, S. Kuoppamäki, A. Hosseini, and G. Skantze, "Between reality and delusion: Challenges of applying large language models to companion robots for open-domain dialogues with older adults," *Autonomous Robots*, vol. 49, no. 1, p. 9, 2025.

[22] E. Hwang *et al.*, "Infusing theory of mind into socially intelligent LLM agents," *arXiv preprint arXiv:2509.22887*, 2025.

[23] X. Zhou, H. Zhu, L. Mathur, R. Zhang, H. Yu, Z. Qi, L.-P. Morency, Y. Bisk, D. Fried, G. Neubig, *et al.*, "SOTOPIA: Interactive evaluation for social intelligence in language agents," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2024.

[24] H. Jiang, X. Zhang, X. Cao, C. Breazeal, D. Roy, and J. Kabbara, "Personallm: Investigating the ability of large language models to express personality traits," in *Findings of the association for computational linguistics: NAACL 2024*, pp. 3605–3627, 2024.

[25] Y. Zhu *et al.*, "Can LLM "self-report"? exploring the validity of LLM-based self-rating in conversational agents," *arXiv preprint arXiv:2412.00207*, 2024.

[26] B. Irfan, J. Miniota, S. Thunberg, E. Lagerstedt, S. Kuoppamäki, G. Skantze, and A. Pereira, "Human-robot interaction conversational user enjoyment scale (HRI CUES)," *IEEE Transactions on Affective Computing*, 2025.

[27] C. Matuszek, T. Williams, N. DePalma, R. Mead, R. Wen, E. Schneiders, C. Kennington, and A. Bezabih, "Reporting guidelines for large language models in human–robot interaction," *J. Hum.-Robot Interact.*, vol. 15, Jan. 2026.

[28] V. Amrhein, S. Greenland, and B. McShane, "Scientists rise up against statistical significance," *Nature*, vol. 567, no. 7748, pp. 305–307, 2019.

[29] A. Amirova, N. Rakhymbayeva, E. Yadollahi, A. Sandygulova, and W. Johal, "10 years of human-NAO interaction research: A scoping review," *Frontiers in Robotics and AI*, vol. 8, p. 744526, 2021.

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 28492–28518, 2023.

[31] R. Xu *et al.*, "Academically intelligent LLMs are not necessarily socially intelligent," *arXiv preprint arXiv:2403.06591*, 2024.

[32] D. Kulić and E. A. Croft, "Affective state estimation for human-robot interaction," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 991–1000, 2007.

[33] H. Voß, H. Wersing, and S. Kopp, "Addressing data scarcity in multimodal user state recognition by combining semi-supervised and supervised learning," in *Companion Publication of the Int. Conf. on Multimodal Interaction (ICMI)*, 2021.

[34] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, pp. 216–225, 2014.

[35] S. Singh, E. Saber, P. P. Markopoulos, and J. Heard, "Regulating modality utilization within multimodal fusion networks," *Sensors*, vol. 24, no. 18, p. 6054, 2024.