# When Embodiment Backfires: The Impact of AI Chatbot Avatars on Human Perceived Empathy

Guanheng Cen*
gc3023@nyu.edu
New York University Shanghai

Xiang Dai*
xd2192@nyu.edu
New York University Shanghai

Xiaowen Zhu*
xz4052@nyu.edu
New York University Shanghai

João Sedoc
jsedoc@stern.nyu.edu
New York University Stern School of Business

## Abstract

Visually embodied conversational AI agents are often assumed to enhance perceived empathy during human–AI conversations by increasing social presence and emotional engagement. However, it remains unclear whether avatar-based interfaces reliably improve users' perceived empathy in practice. We examine this assumption by comparing avatar-based and text-based chatbots across 186 human–AI interactions while holding the conversational context and prompt consistent. Surprisingly, participants interacting with the text-based chatbot reported significantly higher perceived empathy than those interacting with the avatar-based chatbot ($M_{\text{text}} = 5.25$ vs. $M_{\text{avatar}} = 4.86$, $p = 0.038$). This difference persisted after controlling for individual differences in personality traits, demographic characteristics, and conversation length. These results indicate that visual embodiment does not automatically enhance empathic experience and may instead shift the standards by which users evaluate empathic behavior.

## Keywords

Human–AI Interaction; Embodied Conversational Agents; Avatar-Based Chatbots; Perceived Empathy

## 1 Introduction

Embodied conversational agents (ECAs) are often assumed to enhance empathic experience in human–AI interaction by providing richer social cues, including facial expressions, gaze behavior, and vocal prosody. Prior work under the Computers as Social Actors paradigm demonstrates that users readily attribute social and emotional qualities to computational systems [19]. Building on this insight, research on social presence and embodiment argues that visual anthropomorphism can increase users' sense of copresence and social engagement [12, 21]. These assumptions have motivated the deployment of avatar-based chatbots in emotionally sensitive domains such as healthcare and mental health support [30].

However, empirical findings regarding whether visual embodiment reliably improves users' perceived empathy remain mixed. While some studies suggest that humanlike representations can enhance engagement or warmth, others indicate that increased anthropomorphism may introduce elevated expectations that conversational systems fail to satisfy [10, 18]. Expectancy Violations Theory provides a useful framework for understanding this dynamic, positing that unmet social expectations can lead to negative evaluations of an interaction partner [8].

In addition to interface modality, individual differences and representational cues may further shape how users interpret empathic behavior. Prior work has examined user-perceived empathy in chatbot interactions [16, 20, 27] and highlights the role of subjective interpretation in empathy judgments. Visual characteristics of embodied agents, including demographic and gender-related cues, have also been shown to influence perceived warmth, trust, and social evaluation, though findings remain inconsistent [4, 5, 9]. These mixed results underscore the need to systematically examine embodiment effects on perceived empathy across user populations and avatar representations.

Here, as shown in Figure 1, we conduct a comparison between avatar-based and text-based chatbots that share the same conversational backend and response generation pipeline, allowing us to attempt to isolate the effect of visual embodiment on perceived empathy.[1] Using data from 186 human–AI interactions, we examine overall differences in empathic perception across interface modalities, while accounting for individual differences in personality traits and demographic characteristics. Drawing on Expectancy Violations Theory, we explore whether embodiment introduces heightened affective expectations that, when unmet during the interaction, lead users to evaluate the interaction less favorably.

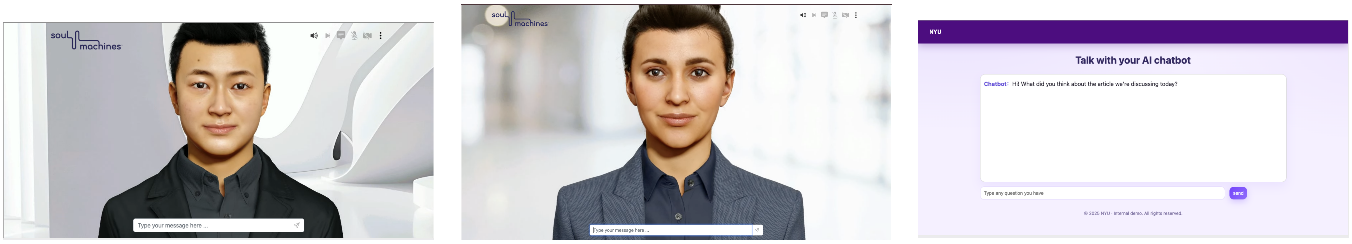
*Research Questions.* This study aims to answer the following research questions:

(1) Does visual embodiment improve perceived empathy and conversation quality compared to text-only interaction?
(2) How do individual differences, such as personality traits and demographic factors, account for variation in users' perceived empathy during interactions with embodied versus text-only chatbots?

This work not only advances the academic understanding of empathy in human–AI dialogue but also lays the groundwork for practical applications of embodied chatbots in emotionally intelligent and socially adaptive communication systems. To examine these questions, we developed two interactive AI chatbot systems and designed a comprehensive experimental and analytical pipeline to support systematic investigation.

*These authors contributed equally to this work.

[1] A demonstration video showcasing interactions with the avatar-based chatbot is available at: https://drive.google.com/file/d/1v6NHywtvV2pYwQpqr7Gruzo9rGJw9tw5/view.




**(a) Avatar-based chatbot (male)** **(b) Avatar-based chatbot (female)** **(c) Text-based chatbot**

**Figure 1: Comparison of chatbot interfaces used in the study: avatar-based (male and female) and text-based conditions.**

*Contributions.* This study makes the following contributions:

(1) **Empirical evidence that visual embodiment can reduce perceived empathy.** We demonstrate that interface modality significantly affects users' empathic evaluations as text-based interactions receive higher perceived empathy scores. This challenges the common assumption that richer social cues automatically improve affective experience.
(2) **Evidence for an anthropomorphism-mediated mechanism.** We find that avatar-based interfaces, counterintuitively, are perceived as less humanlike than text-based interfaces, and that this reduction in anthropomorphism mediates the modality effect on perceived empathy.
(3) **A controlled experimental dataset of human–AI interactions.** We collect and release 186 paired dialogue logs and post-interaction survey responses across avatar and text conditions, including perceived empathy, conversation quality, trust, personality traits, and demographics, enabling future replication and extension.

## 2 Background & Related Work

Our work directly builds upon the experimental paradigm of Liu et al. [16], who find that while the empathetic language of chatbots and humans is equivalent when analyzed computationally, users nonetheless perceive chatbots as less empathetic than humans. This disconnect between empathetic language generation and user perception directly motivated our investigation into whether visual embodiment might help close this gap.

### 2.1 Avatar-based and Text-based Chatbots

Our research intersects with literature on embodied conversational agents, social presence theory, and human–AI interaction. The Computers as Social Actors (CASA) framework [19] establishes that users apply social rules to computer interfaces, extending to avatar-based chatbots where anthropomorphic cues encourage users to treat the system as a social actor [12]. Media richness theory [28] and digital embodiment work [21] suggest that richer social cues enhance perceived presence. Recent work extends to LLM-driven agents: Lee et al. [14] demonstrate that GPT-4 responses are perceived as more emotionally attuned when delivered by an expressive avatar, while digital health assistant studies report increased engagement with embodied agents [30]. However, embodiment effects are not uniformly positive. Imperfect facial expressions can trigger uncanny-valley reactions [18].

Several critical limitations constrain clear conclusions about embodiment's impact on perceived empathy. First, limited systematic variation in avatar characteristics: prior studies typically employ a single avatar identity or do not systematically vary characteristics such as gender, limiting understanding of whether embodiment effects generalize across different presentations. Second, inadequate control of linguistic content: most comparisons do not sufficiently align conversational behavior, leaving open whether perceived empathy differences reflect interface cues versus differences in what the system actually says. This is especially problematic for LLM-driven systems, where prompt engineering, model parameters, or conversation history could introduce systematic differences beyond presentation modality. Third, narrow focus on usability outcomes: applied evaluations often emphasize usability and engagement metrics [30], but less frequently examine whether perceived empathy differences reflect interface cues versus underlying textual responses, or how individual differences moderate embodiment effects.

Our study addresses these limitations through a carefully controlled experimental design. We construct two chatbot systems that differ only in presentation modality—avatar-based versus text-only—while sharing identical prompts for response generation and conversation management pipeline. This enables a direct test of whether embodiment itself changes perceived empathy beyond language content. We further incorporate both male-presenting and female-presenting avatars to examine whether embodiment effects appear consistent across different avatar presentations.

### 2.2 Perceived Empathy and Conversation Quality in Human–AI Interaction

Perceived empathy connects to research on how users evaluate emotional understanding in human–AI conversations. Prior work emphasizes that empathy is fundamentally experiential: users base judgments on subjective feelings of being heard rather than explicit empathic markers [6, 27]. Studies show that self-reported empathy aligns with perceived attentiveness even when textual content is held constant [16], and people can perceive high empathy even when responses are produced by LLMs [14]. Perceived conversation quality reflects users' judgments of coherence, naturalness, and responsiveness. LLM-generated responses often receive high quality ratings due to clarity and stylistic fluency, even when emotional depth is limited [1], while empathic phrasing can improve perceived flow [31]. Computational approaches to empathy modeling have advanced through shared tasks [11, 24] and methods for text-based mental health support [26]. However, Liu et al. [16] find that conversations can achieve high quality ratings while receiving lower empathy ratings, suggesting users evaluate "how well the conversation went" and "how emotionally understood they felt" through different cognitive processes.

One limitation that constrains understanding of how embodiment shapes perceived empathy and conversation quality is conceptual conflation. Empathy and conversation quality are sometimes treated as interchangeable outcomes, despite evidence they can dissociate. Our research addresses this limitation by measuring perceived empathy and conversation quality as distinct constructs using validated scales, enabling independent testing of embodiment effects. We test whether modality-driven differences persist when linguistic content is held constant by design.

## 3 Methods

We adopt the experimental design of Liu et al., which draws on the Empathic Conversations dataset and the WASSA 2023 and 2024 shared task corpora [2, 11, 23]. In this paradigm, crowdworkers read a negative news article selected to elicit empathic reactions [7] and then engage in a multi-turn conversation, with pre- and post-conversation surveys administered at each end. We follow this design closely, retaining the same news articles.

To capture participants' subjective perceptions of chatbot interactions, we administered post-conversation surveys using three well-established psychological measurement instruments.

### 3.1 Experimental Design

We recruited participants through Prolific and completed a Qualtrics survey that embedded either the avatar-based or text-based chatbot interface, along with pre- and post-conversation measures of demographic characteristics and personality traits. Participants engaged in a multi-turn conversation through either a text only frontend or avatar frontend where both connected to the same LLM backend for response generation. The avatar-based condition was implemented using the Soul Machines Digital Human platform. The selected avatars provided a clear visual embodiment of the conversational agent while maintaining experimental control over interface modality (see Figure 2and Figure 3).

### 3.2 Survey Design

To complement our computational modeling and deepen our understanding of participants' subjective reactions, we incorporated three established psychological measurement instruments: the Godspeed Questionnaire Series (GQS), the 20-item Big Five Personality Inventory (BFI-20), and SPUR Social Presence Items.
These instruments capture perceptions of embodied agents, individual personality traits, and perceived social presence, respectively. Detailed descriptions of each questionnaire and their constructs are provided in Appendix B.

#### Godspeed Questionnaire Series (GQS)

We use the GQS to measure participants' perceptions of embodied agents, including anthropomorphism, animacy, likeability, intelligence, and perceived safety. These dimensions are widely used to evaluate human–agent interaction and provide a structured way to capture subjective impressions.

#### 20-Item Big Five Personality Inventory

To control for individual differences, we include the BFI-20 to capture personality traits such as Extraversion and Agreeableness, which are particularly relevant to social interaction and empathy perception.

#### SPUR Social Presence Items

We use SPUR-based items to measure perceived social presence in human–AI interaction, focusing on perceived conversational quality, empathy, and emotional understanding.

## 4 Results

We report findings from 186 participants across three experimental conditions. This section describes the data collection process and presents statistical analyses of perceived empathy differences.

### 4.1 Data

Similar to the setting in the empathic conversation [11, 16], we asked crowdworkers read a negative news article that evokes empathy and then interact with a conversational AI system about this news article. Specifically, participants were randomly assigned to interact with either a text-based chatbot or a male- or female-presenting avatar chatbot, engaged in a short multi-turn conversation, and then completed a survey assessing perceived empathy, conversation quality, trust, personality traits, and demographics. These human-AI interactions were collected through Prolific, and after removing responses from incomplete surveys or very short interactions (fewer than six turns), the final dataset comprised 186 valid conversations: 93 in the text condition, 49 in the male-avatar condition, and 44 in the female-avatar condition. According to the demographics information summarized in Table 1, participants ranged in age from 19 to 80 years (M=43.4, SD=11.8). The sample was approximately balanced by sex (51.1% male, 48.9% female) and predominantly White (69.4%), with representation from Asian (13.4%), Black (10.2%), Mixed (3.8%), and Other (3.2%) ethnic groups. Demographic characteristics were generally well-distributed across conditions, with some variation in subgroup sizes due to the stratified sampling approach.

### 4.2 Statistical Analysis

In the statistical analysis of perceived empathy scores were compared across the two modalities using independent-samples t-tests and OLS regression models controlling for individual differences.

| Ethnicity | SRG | Condition | | |
|---|---|---|---|---|
| | | Text<br>*n*, Age M (SD) | Male Avatar<br>*n*, Age M (SD) | Female Avatar<br>*n*, Age M (SD) |
| White | Male | 33, 49.8 (10.7) | 15, 41.7 (12.9) | 17, 43.5 (12.2) |
| | Female | 32, 41.1 (9.7) | 18, 45.0 (10.4) | 14, 50.4 (14.7) |
| Asian | Male | 8, 35.1 (8.9) | 3, 30.7 (9.3) | 4, 34.0 (2.2) |
| | Female | 5, 45.2 (9.3) | 3, 38.3 (6.8) | 2, 46.5 (19.1) |
| Black | Male | 4, 46.5 (13.2) | 4, 36.0 (7.3) | 2, 43.5 (9.2) |
| | Female | 7, 38.4 (14.9) | 2, 46.0 (11.3) | — |
| Mixed | Male | 1, 44.0 | 1, 39.0 | — |
| | Female | 1, 46.0 | 2, 43.0 (2.8) | 2, 28.5 (13.4) |
| Other | Male | 1, 23.0 | — | 2, 40.5 (7.8) |
| | Female | 1, 28.0 | 1, 63.0 | 1, 43.0 |

**Table 1: Participant Demographics by Condition, Ethnicity, and Self-reported gender; SRG = Self-reported Gender; M: Mean, SD: Standard Deviation.**

## T-test Results (SPUR)

We conducted independent-samples t-tests to compare perceived empathy, perceived conversation quality, and perceived trust between avatar-based and text-based conditions. These three measures—representing distinct facets of users' felt connection and social engagement—were operationalized using items derived from the SPUR social presence framework [17]. Specifically, perceived empathy was assessed using the EMP composite (computed as the mean of EMP_1 through EMP_5). The analysis revealed a statistically significant difference between conditions, with the text-based chatbot receiving higher perceived empathy scores than the avatar-based chatbot ($t = -2.09$, $p = 0.038$). The text-based condition showed a higher mean empathy score ($M = 5.25$, $SD = 1.24$) compared to the avatar-based condition ($M = 4.86$, $SD = 1.31$). We then examined perceived conversation quality, calculated as the mean of 14 PCQ items. A similar independent-samples t-test indicated a statistically significant difference between conditions ($t = -2.35$, $p = 0.020$), with the text-based condition ($M = 5.85$, $SD = 0.83$) receiving higher scores than the avatar-based condition ($M = 5.55$, $SD = 0.87$). Finally, perceived trust was assessed using the TRUST composite (computed as the mean of TRUST_1 through TRUST_5). The results again showed a statistically significant difference between conditions ($t = -2.07$, $p = 0.039$), with the text-based condition showing higher trust ($M = 5.34$, $SD = 1.16$) than the avatar-based condition ($M = 4.97$, $SD = 1.26$).

## Descriptive Subgroup Patterns

Following the overall comparison, we conducted exploratory analyses to examine whether patterns of perceived empathy varied across participant characteristics and avatar presentation types, including comparisons across ethnicity groups within each condition and between male-presenting and female-presenting avatars. Although mean scores varied slightly across subgroups, the overall pattern of higher empathy ratings in the text-based condition remained generally consistent. Specific descriptive statistics are summarized in Tables 2 and 3.

| Ethnicity Group | Avatar | | Text | |
|---|---|---|---|---|
| | *M* (*SD*) | *n* | *M* (*SD*) | *n* |
| White | 4.83 (1.25) | 64 | 5.09 (1.24) | 65 |
| Asian | 5.25 (1.14) | 12 | 5.60 (1.41) | 13 |
| Black | 4.93 (1.59) | 8 | 5.93 (0.79) | 11 |

**Table 2: Descriptive statistics of Perceived Empathy (EMP Composite) across ethnicity groups; M: Mean, SD: Standard Deviation, n: Sample Size.**

| Condition | *M* (*SD*) | *n* |
|---|---|---|
| Male-like Avatar | 4.91 (1.37) | 49 |
| Female-like Avatar | 4.80 (1.26) | 44 |

**Table 3: Perceived Empathy (EMP Composite) by avatar gender presentation; M: Mean, SD: Standard Deviation.**

## Regression Model

To assess whether the interface modality predicts perceived empathy while accounting for individual demographic characteristics and conversation behavior, we fit an ordinary least squares (OLS) regression model with EMP Composite as a dependent variable. The primary predictor was condition_avatar, coded as 1 for avatar-based interactions and 0 for text-based interactions. The control variables included self-reported gender, ethnicity and total_tokens, where total_tokens was calculated as the total number of tokens in the full conversation between each participant and chatbot. Following the findings in Section 3.3, extraversion and agreeableness were also selected for inclusion in the model as personality control variables because they are particularly relevant to interpersonal perception in our research context. Both were entered as continuous variables. Composite scores for extraversion and agreeableness were computed following established scoring procedures, with reverse-coded items handled accordingly [29].

Table 4 reports the results of the OLS regression predicting perceived empathy (EMP Composite) while controlling for individual demographic characteristics and conversation length. Consistent with the t-test results, interface modality remained a statistically significant predictor of perceived empathy after accounting for personality category, sex, ethnicity, and total token count. Specifically, the coefficient for condition_avatar was negative and significant ($\beta = -0.438$, $p = 0.018$), indicating that avatar-based interactions were associated with lower perceived empathy scores compared to text-based interactions, holding other variables constant. In contrast, total_tokens did not exhibit a significant association with perceived empathy ($p = 0.327$). Together, these findings provide converging evidence that interface modality plays a meaningful role in shaping users' empathy perceptions, reinforcing the pattern observed in the preceding t-test analyses.

| DV: | Perceived Empathy | | | |
|---|---|---|---|---|
| Models: | (1) | (2) | (3) | (4) |
| condition_avatar | $-0.391^{*}$ | $-0.350^{*}$ | $-0.422^{*}$ | $-0.438^{*}$ |
| | (0.187) | (0.187) | (0.183) | (0.184) |
| *Control Variables* | | | | |
| Demographics | NO | YES | YES | YES |
| Personality | NO | NO | YES | YES |
| Total Conv. Tokens | NO | NO | NO | YES |
| Observations | 186 | 186 | 183 | 183 |
| R-squared | 0.018 | 0.068 | 0.153 | 0.158 |

**Table 4: Main Effects of condition_avatar (dummy variable; avatar = 1, text = 0) on Perceived Empathy (EMP Composite); Standard errors are reported in parentheses.** $^{*}p < 0.05$.

| DV: | Perceived Empathy | | Anthropomorphism |
|---|---|---|---|
| Models: | (1) | (2) | (3) |
| condition_avatar | $-0.438^{*}$ | $0.288^{*}$ | $-0.809^{*}$ |
| | (0.184) | (0.131) | (0.154) |
| anthro_score | – | $0.898^{*}$ | – |
| | – | (0.060) | – |

**Table 5: Models for exploring mediation effect of anthropomorphism on modality-empathy relationship. Model (1) reproduces the baseline Model (4) from Table 4. Model (2) adds composite anthropomorphism score as an additional predictor; Standard errors are reported in parentheses.** $^{*}p < 0.05$.

## Analysis of Anthropomorphism

To further investigate the mechanism underlying the modality effect, we analyzed the Godspeed Questionnaire Series (GQS), which assesses perceived qualities of the agent across five subscales: Anthropomorphism (extent to which the agent is perceived as humanlike), Animacy, Likeability, Intelligence, and Safety. For each subscale we computed a composite score by averaging the corresponding items and compared avatar-based and text-based conditions using a between-group $t$-test. Participants in the text-based condition reported higher scores than those in the avatar-based condition on Anthropomorphism ($M = 3.52$, $SD = 1.00$ vs. $M = 2.76$, $SD = 1.08$; $t = -5.01$, $p < .001$), Animacy ($M = 3.68$, $SD = 1.02$ vs. $M = 3.22$, $SD = 0.94$; $t = -3.18$, $p = .002$), and Likeability ($M = 4.23$, $SD = 0.77$ vs. $M = 3.98$, $SD = 0.77$; $t = -2.20$, $p = .029$). The difference was marginal for Intelligence (text $M = 4.24$, $SD = 0.78$; avatar $M = 4.05$, $SD = 0.73$; $t = -1.75$, $p = .083$), and there was no significant difference for Safety ($p = 1.00$). In prior work on human–agent interaction, perceived humanlikeness has been shown to foster empathy toward agents [15], so we focus on the Anthropomorphism subscale as the most theoretically relevant to explain modality-related differences in perceived empathy. The between-group $t$-test reported above shows that the text-based condition elicited higher anthropomorphism scores than the avatar-based condition, mirroring the pattern observed for perceived empathy.

Additionally, we examined whether anthropomorphism mediates the relationship between modality and perceived empathy using OLS regressions that extend the baseline model in Table 4, Model 4. We estimated the additional specifications summarized in Table 5 using the same full set of control variables, and all regression results are reported in that table. In Model 3, where anthropomorphism is the dependent variable, condition_avatar significantly predicted anthropomorphism scores ($\beta = -0.81$, $p < .001$), indicating that participants in the text condition perceived the agent as more humanlike than those in the avatar condition. In Model 2, which adds anthropomorphism as a predictor of perceived empathy, anthropomorphism was a strong positive correlate of empathy ($\beta = 0.90$, $p < .001$), and the coefficient for condition_avatar changed sign and became positive ($\beta = 0.29$, $p = .029$). These patterns are consistent with a negative indirect effect of modality on perceived empathy operating through anthropomorphism, where avatar condition leads to lower anthropomorphism which in turn predicts lower empathy.

# 5 Discussion

The results suggest that interface presentation can shape how users assess empathic qualities in human–AI interaction, while maintaining the same conversational context and prompt. Rather than enhancing perceived empathy, visual embodiment appears to influence the evaluative frame through which empathic experience is interpreted.

Expectancy Violations Theory offers a plausible interpretive account for this pattern. Visual embodiment introduces salient social and humanlike cues that may contribute to how users form expectations prior to and during interaction. When an interface includes an avatar, users may implicitly apply more human-centered or demanding criteria when judging empathic qualities. In contrast, a text-based interface may cue a different expectation frame, leading users to evaluate the same interaction using less stringent standards. At the same time, alternative mechanisms may also contribute to the observed pattern. Avatar-based interfaces introduce additional visual cues and processing demands that may shift users' attention toward the interface presentation itself rather than the conversational content. In addition, Schlesener et al. [25] distinguish between embodied anthropomorphic form (EA), which refers to the visual presentation of an agent, and behavioral anthropomorphic form (BA), which refers to conversational behaviors that demonstrate understanding of users' mental states. Their results suggest that users' perceptions of conversational agents are shaped by the interaction between EA and BA rather than embodiment alone. In our context, introducing an avatar primarily adds EA without necessarily increasing BA, which may contribute to lower empathic evaluations when the conversational behavior does not match the expectations implied by the embodied interface.

Several limitations of the present study should be acknowledged. First, although the sample size was sufficient to detect overall differences in perceived empathy, it may limit the detection of smaller moderation effects across demographic subgroups. Second, the avatar representations examined were restricted to male-presenting and female-presenting forms, and other visual dimensions—such as age cues, ethnicity cues, facial expressiveness, or degrees of realism—were not systematically varied. Third, the study focused exclusively on users' subjective evaluations of empathic experience, which constrains the scope of inference to perception rather than behavioral or linguistic properties of the interaction.

## 6 Conclusion

This study provides a controlled empirical examination of how visual embodiment influences users' perceived empathy in human–AI interaction, as reflected in subjective evaluations. Across 186 interactions, we find a consistent main effect of interface modality: participants interacting with a text-based chatbot reported significantly higher perceived empathy than those interacting with an avatar-based chatbot. This effect was observed in both independent-samples comparisons and regression analyses, and remained significant after controlling for conversation length, personality category, sex, and ethnicity. Rather than enhancing perceived empathy, the presence of an embodied avatar was associated with a reduction in users' empathic evaluations.

These findings suggest that visual embodiment does not automatically improve empathic experience in this specific implementation and may instead shift the evaluative standards users apply during interaction. Consistent with Expectancy Violations Theory, the presence of an avatar may elevate users' affective expectations, thereby prompting more stringent evaluations of the same conversational responses compared to text-only interaction. Other mechanisms such as differences in attentional demands, perceptual processing, or the alignment between visual embodiment and conversational behavior may also contribute to how empathic qualities are evaluated. From a design perspective, this result implies that introducing visual embodiment in empathetic AI systems should be approached cautiously, as richer social cues may alter evaluative standards rather than automatically improving users' perceived empathic experience.

## Acknowledgments

We would like to thank our reviewers for their valuable comments. This work was supported by funding from Wellcome and supported by Neuromatch, MEXA grant and NYU's Dean's Undergraduate Research Fund (DURF).

## References

[1] Joseph W. Ayers, Adam Poliak, William Hersh, Aaron Goodman, and Davey M. Smith. 2023. Comparing Physician and ChatGPT Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine* 183, 6 (2023), 595–602. doi:10.1001/jamainternmed.2023.1838

[2] Valentin Barriere, Joao Sedoc, Shabnam Tafreshi, and Salvatore Giorgi. 2023. Findings of WASSA 2023 Shared Task on Empathy, Emotion and Personality Detection in Conversation and Reactions to News Articles. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis*. Association for Computational Linguistics, Toronto, Canada. Available at https://aclanthology.org/2023.wassa-1.44.pdf.

[3] Christoph Bartneck. 2023. Godspeed Questionnaire Series: Translations and Usage. *ResearchGate Preprint / Chapter* (2023). https://www.researchgate.net/publication/367970840_Godspeed_Questionnaire_Series_Translations_and_Usage Accessed: YYYY-MM-DD.

[4] Christoph Bartneck, Denisa Kulić, Elizabeth Croft, and Sherif Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1, 1 (2009), 71–81. doi:10.1007/s12369-008-0001-3

[5] Christoph Bartneck, Kumar Yogeeswaran, Qi Min Ser, Graeme Woodward, Robert Sparrow, Siheng Wang, and Friederike Eyssel. 2018. Robots and Racism. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI)* 7 (2018), 196–204. doi:10.1145/3171221.3171260

[6] Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and Maintaining Long-Term Human–Computer Relationships. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. 197–204. doi:10.1145/1054972.1054999

[7] Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4758–4765.

[8] Judee K. Burgoon. 2015. Expectancy Violations Theory. *The International Encyclopedia of Interpersonal Communication* (2015). doi:10.1002/9781118540190.wbeic102 Accessed: 2025-12-11.

[9] Michelle Cohn, Natalie Friedman, and James Smith. 2019. Age Cues in Digital Avatars Influence Perceived Trust and Sociability. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. doi:10.1145/3290607.3312761

[10] Michael Curry and Verena Rieser. 2023. Overtrust in Human-like AI Systems: Ethical and Psychological Implications. *AI and Ethics* (2023).

[11] Salvatore Giorgi, Joao Sedoc, Valentin Barriere, and Shabnam Tafreshi. 2024. Findings of WASSA 2024 Shared Task on Empathy and Personality Detection in Interactions. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis*. Association for Computational Linguistics, Bangkok, Thailand. Available at https://aclanthology.org/2024.wassa-1.30.pdf.

[12] Younghwa Go and S. Shyam Sundar. 2019. Humanizing Chatbots: The Effects of Visual, Identity and Conversational Cues on Humanness Perceptions. *Computers in Human Behavior* 97 (2019), 304–316. doi:10.1016/j.chb.2019.01.020

[13] Lewis R. Goldberg. 1992. The development of markers for the Big-Five factor structure. *Psychological Assessment* 4, 1 (1992), 26–42.

[14] Yoon Kyung Lee, Jina Suh, Hongli Zhan, Junyi Jessy Li, and Desmond C. Ong. 2024. Large Language Models Produce Responses Perceived To Be Empathic. *arXiv preprint arXiv:2403.18148* (2024).

[15] Jingshu Li, Zicheng Zhu, Renwen Zhang, and Yi-Chieh Lee. 2025. Exploring the Effects of Chatbot Anthropomorphism and Human Empathy on Human Prosocial Behavior Toward Chatbots. *Proceedings of the ACM on Human-Computer Interaction* 9, 7, Article CSCW415 (November 2025). doi:10.1145/3757596

[16] Tingting Liu, Salvatore Giorgi, Ankit Aich, Allison Lahnala, Brenda Curtis, Lyle Ungar, and Joao Sedoc. 2025. The Illusion of Empathy: How AI Chatbots Shape Conversation Perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence, Philadelphia, PA. Available at https://arxiv.org/abs/2411.12877.

[17] Patrick R. Lowenthal. 2009. Social Presence. *Encyclopedia of Distance Learning (2nd ed.)* (2009). https://patricklowenthal.com/publications/2009/2009--Social-Presence.pdf Accessed: 2025-12-11.

[18] Maya B. Mathur and David B. Reichling. 2015. Navigating a Social World with Robot Partners: A Quantitative Cartography of the Uncanny Valley. *Computers in Human Behavior* 146 (2015), 22–32. doi:10.1016/j.cognition.2015.09.008

[19] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1 (2000), 81–103. doi:10.1111/0022-4537.00153

[20] Trang T. Nguyen, David C. Hernandez, Erin Beneteau, Q. Vera Liao, and Wai-Tat Fu. 2022. User-Perceived Empathy in Chatbot Interactions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–16. doi:10.1145/3491102.3517691

[21] Kristine L. Nowak and Frank Biocca. 2003. The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence. *Presence: Teleoperators and Virtual Environments* 12, 5 (2003), 481–494. doi:10.1162/105474603322761289

[22] C. S. Oh, Jeremy N. Bailenson, and Gregory F. Welch. 2018. A Systematic Review of Social Presence: Definition, Antecedents, and Implications. *Frontiers in Robotics and AI* 5 (2018), 114. doi:10.3389/frobt.2018.00114

[23] Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic Conversations: A Multi-level Dataset of Contextualized Conversations. *arXiv preprint arXiv:2205.12698* (2022). https://arxiv.org/abs/2205.12698 Available at https://arxiv.org/pdf/2205.12698.pdf.

[24] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5370–5381. https://aclanthology.org/P19-1534/

[25] Elizabeth A. Schlesener, Marcin Ziolkowski, Sai-Keung Wong, Brent Westmoreland, and Sabarish V. Babu. 2026. "Am I Understood?": How the Interplay between Embodiment and Theory of Mind Behavior Affects LLM-Based Conversational Agents on Perceived Trust, Anthropomorphism, Presence, Usability, and User Experience. *ACM Transactions on Interactive Intelligent Systems* 16, 2 (2026). doi:10.1145/3774779

[26] Eva Sharma, Daniel Preoţiuc-Pietro, Matthew Matero, and Lyle Ungar. 2020. A Computational Approach to Understanding Empathy in Text-based Mental

Health Support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 5263–5276. https://aclanthology.org/2020.emnlp-main.425/
[27] Lijiang Shen. 2010. On a Scale of State Empathy During Message Processing. *Western Journal of Communication* 74, 5 (2010), 504–524. doi:10.1080/10570314.2010.512278
[28] John Short, E. Williams, and B. Christie. 1976. *The Social Psychology of Telecommunications*. John Wiley & Sons.
[29] Ewa Skimina, Włodzimierz Strus, Ewa Topolewska-Siedzik, Tomasz Rowiński, and Jan Cieciuch. 2014. The short IPIP-BFM-20 questionnaire for measuring the Big Five. *Roczniki Psychologiczne* 17, 2 (2014), 385–402.
[30] Almira Osmanovic Thunström, Hanne Krage Carlsen, Lilas Ali, Tomas Larson, Andreas Hellström, and Steinn Steingrimsson. 2024. Usability Comparison Among Healthy Participants of an Anthropomorphic Digital Human and a Text-Based Chatbot as a Responder to Questions on Mental Health: Randomized Controlled Trial. *JMIR Human Factors* (2024). doi:10.2196/54581
[31] Zhichao Xu and Jiepu Jiang. 2024. Multi-dimensional Evaluation of Empathetic Dialog Responses. *arXiv preprint arXiv:2402.11409* (2024).

# A Chatbot Interface Implementation

This section describes the implementation of two chatbot interfaces developed for the comparative study. Both systems are designed to facilitate empathetic conversations about news articles with users, sharing identical core functionality: (1) session management and parameter extraction from URL query strings, (2) conversation context construction incorporating article background and historical dialogue, (3) AI response generation through the OpenAI API using identical prompts and parameters, and (4) conversation persistence to MongoDB using the same document schema.

The fundamental difference between the two implementations lies in their interaction modalities. The *Avatar-Based Chatbot* provides an embodied conversational experience through Soul Machines Digital Person technology, enabling multimodal interaction via voice and video. Users interact with a virtual avatar that renders in real-time, with facial animations synchronized to speech output. In contrast, the *Text-Based Chatbot* employs a traditional text-only interface, where users type messages and receive text responses in a standard chat window.

Both systems converge at the backend processing layer, where identical conversation logic is used to minimize differences in AI behavior across conditions. This architectural consistency allows observed differences in user experience to be attributed to interface modality—embodied avatar versus text-only—rather than to variation in system functionality.

## A.1 Frontend Implementation

The Avatar-Based Chatbot frontend employs a component-based architecture organized into modular layers, as illustrated in Figure 2 and Figure 3. The main conversation interface component orchestrates the overall interaction flow, coordinating between Digital Person rendering components, user interaction components, and content display components (for presenting conversation history and supplementary information). A centralized state management infrastructure maintains connection status, media permissions, conversation history, and Digital Person response states.

The frontend establishes a WebSocket connection to the Soul Machines platform upon initialization, enabling real-time bidirectional communication for audio and video streaming. User interactions are captured through either voice input (streamed to the platform for speech-to-text processing) or text input (a fallback mode), with the Digital Person video rendered in real-time and synchronized

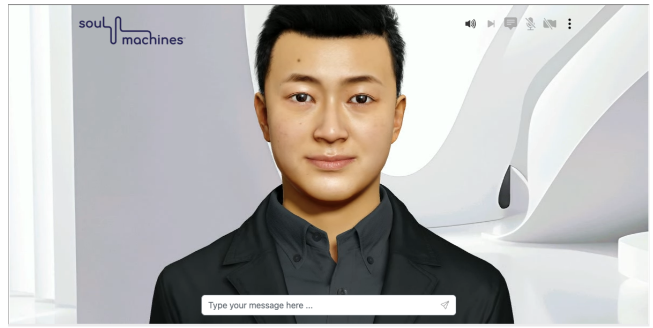



**Figure 2: Avatar-Based Chatbot frontend interface (male-presenting avatar)**

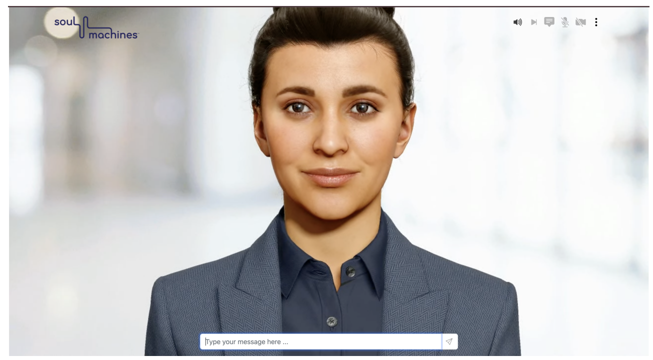



**Figure 3: Avatar-Based Chatbot frontend interface (female-presenting avatar)**

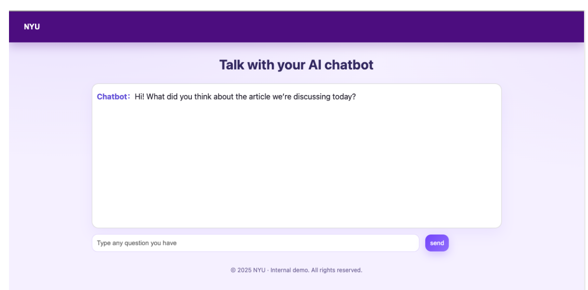



**Figure 4: Text-Based Chatbot frontend interface**

with audio output and facial animations. The connection lifecycle is managed through component lifecycle hooks, ensuring proper resource cleanup.

The Text-Based Chatbot frontend follows a simplified single-page architecture, as shown in Figure 4. The interface consists of three core elements: (1) a chat history display area, (2) a text input field, and (3) a send button. This lightweight structure eliminates the complexity of component hierarchies and state management frameworks while maintaining core functionality.

User messages are transmitted directly to the backend via HTTP POST requests, with responses immediately rendered in the chat interface. Both frontend implementations share identical session management logic and parameter extraction strategies: experimental parameters are extracted from URL query strings and persisted in browser sessionStorage to maintain continuity across page refreshes, ensuring experimental consistency.

## A.2 Backend Implementation

Both chatbot systems implement identical conversation processing logic at the backend layer, ensuring behavioral consistency. The key architectural difference lies in how the backend integrates with the frontend and AI services, as illustrated in Figure 5.

*Avatar-Based Backend Architecture.* The Avatar-Based Chatbot backend is implemented as a Soul Machines Skill, adhering to the Skills API specification. The Skill exposes four standard endpoints that integrate with the Soul Machines platform lifecycle: `POST /init` (project initialization), `POST /session` (session management and parameter extraction), `POST /execute` (conversation execution), and `DELETE /delete/{project_id}` (resource cleanup). The platform acts as an intermediary, invoking these Skill endpoints when processing user interactions, and the Skill communicates directly with the OpenAI API for response generation.

*Text-Based Backend Architecture.* The Text-Based Chatbot backend is implemented as a standard FastAPI application with two endpoints: `GET /` (serves the HTML interface) and `POST /chat` (processes user messages). This architecture eliminates the intermediate platform layer, establishing direct HTTP communication between the frontend and backend. The backend communicates directly with the OpenAI API, mirroring the same conversation processing logic as the Avatar-Based implementation.

*Shared Conversation Processing Logic.* Both backend implementations employ identical conversation processing logic. The process involves: (1) extracting user input and session identifiers, (2) building a comprehensive conversation context that includes system prompts (empathy-focused behavioral guidelines), article background information (limited to 10,000 characters), and conversation history retrieved from the database, (3) invoking the OpenAI API with identical parameters (GPT-4.1-nano, temperature 0.7) to generate responses, (4) managing completion code logic to ensure a minimum number of interaction turns for experimental control, and (5) persisting conversation data to MongoDB using the same document schema.

To balance context preservation with API token constraints, both systems employ a sliding window approach: for conversations exceeding 50 turns, the system preserves the first 5 turns (capturing important opening context) and the most recent 45 turns (maintaining recent continuity). Parameter extraction strategies differ slightly between implementations: the Avatar-Based system extracts parameters from nested context objects provided by the Soul Machines platform, while the Text-Based system extracts from HTTP request bodies and URL query parameters. However, both employ hierarchical fallback mechanisms to support reliable parameter retrieval.

## A.3 Data Persistence and Schema

All conversation data is persisted in MongoDB Atlas using a standardized document schema shared by both systems. Each conversation record contains: session identifier, article identifier, response identifier (for experimental tracking), turn number, user input text, bot response text, and timestamps for both user input and bot response. This unified schema enables efficient conversation history retrieval for context building during conversations and comprehensive post-experiment analysis.

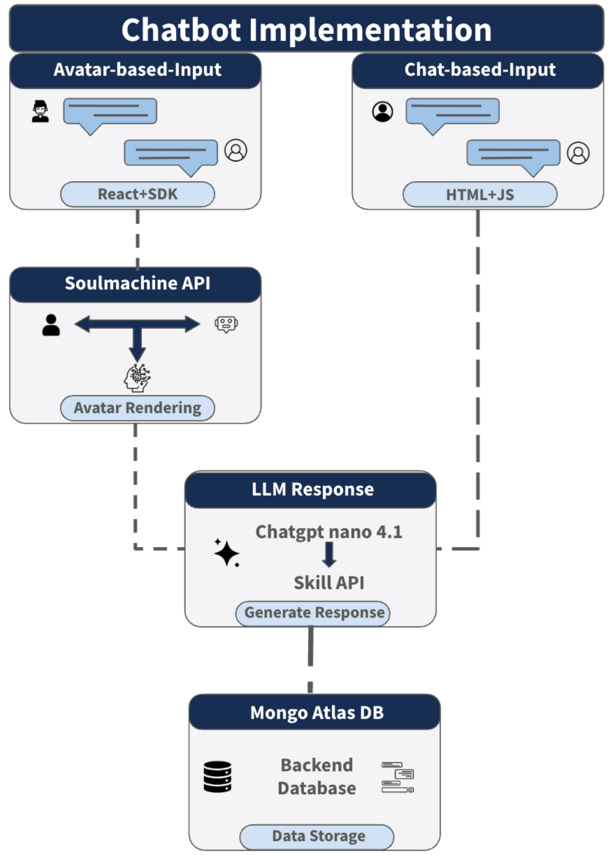



**Figure 5: Chatbot Implementation Architecture**

# B Measurement Instruments

## B.1 Godspeed Questionnaire Series (GQS)

The Godspeed Questionnaire Series was developed by Bartneck et al. (2009) to provide a standardized, cross-study measure of how humans perceive artificial agents [4]. Prior to GQS, the field lacked consistent metrics for evaluating subjective impressions of robots or virtual entities. Since its publication, Godspeed has become one of the most widely used and frequently cited evaluation tools in HCI and social robotics, with thousands of citations across domains such as embodied conversational agents, VR avatars, and service robots [3].

The questionnaire assesses several dimensions central to users' interpretations of embodied AI systems: Anthropomorphism (the extent to which the agent seems human-like), Animacy (the degree to which the agent appears alive or biologically alive), Likeability (warmth, friendliness, and emotional valence), Perceived Intelligence (users' assessment of the agent's reasoning and competence), and Perceived Safety (the degree to which the agent appears non-threatening and comfortable to interact with) [3]. These constructs help us evaluate whether avatar-based embodiments influence participants' impressions in ways that could shape perceived empathy. In this sense, Godspeed provides a validated foundation for determining whether changes in embodiment shift users' interpersonal judgments independently of linguistic content.

## B.2 20-Item Big Five Personality Inventory

To account for stable individual differences, we administered the 20-item Big Five personality measure from the International Personality Item Pool (IPIP-20) [13]. The Big Five framework—Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism—constitute the dominant framework in contemporary personality psychology, supported by extensive empirical evidence across cultures and contexts [13]. This inventory allows us to capture trait-level differences that may shape how participants interpret social cues and attribute empathy during interaction.

Although the instrument captures five personality dimensions, our analysis focused on Extraversion and Agreeableness, as these traits are most closely related to social orientation and interpersonal sensitivity in our research settings. Extraversion encompasses sociability, social engagement, and interpersonal expressiveness, while Agreeableness reflects warmth, empathy, prosocial orientation, and interpersonal sensitivity [13]. Both traits have been consistently identified as socially relevant across research traditions examining interpersonal perception and human–agent interaction, and are commonly associated with social orientation and interpersonal sensitivity in studies of empathy and social presence [23].

Including personality measures strengthens the internal validity of our analysis by enabling us to control for psychological variation that could otherwise confound the relationship between embodiment and perceived empathy. For example, individuals higher in Agreeableness may be more inclined to attribute warmth to the agent, while those higher in Extraversion may exhibit different patterns of social responsiveness when engaging with embodied versus text-based interfaces.

## B.3 SPUR Social Presence Items

To examine the psychological mechanism underlying the relationship between embodiment and perceived empathy, we incorporate measurement items from the SPUR framework, which is grounded in the foundational Social Presence Theory introduced by Short, Williams, and Christie (1976) [17]. Social presence refers to the subjective sense that one is interacting with an intentional, emotionally expressive, and "real" social entity. Over the decades, the concept has been expanded within modern HCI contexts including VR/AR agents, virtual humans, conversational AI, and telecommunication systems [22]. SPUR-style items are now widely recognized as standard measures for quantifying users' felt connection, emotional engagement, and perceived warmth in mediated interactions [22].

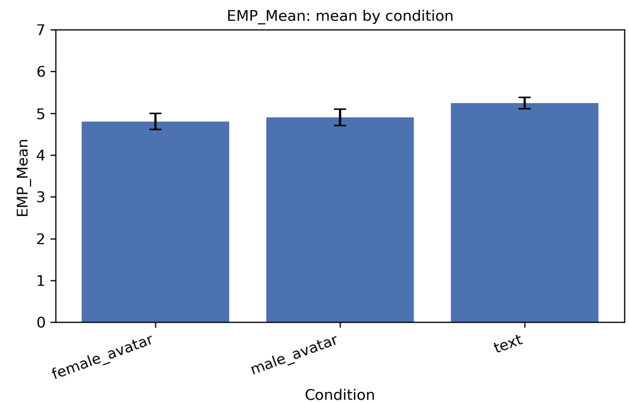



**Figure 6: Mean perceived empathy (EMP composite) by condition. Corresponds to Table 5, Model (1).**

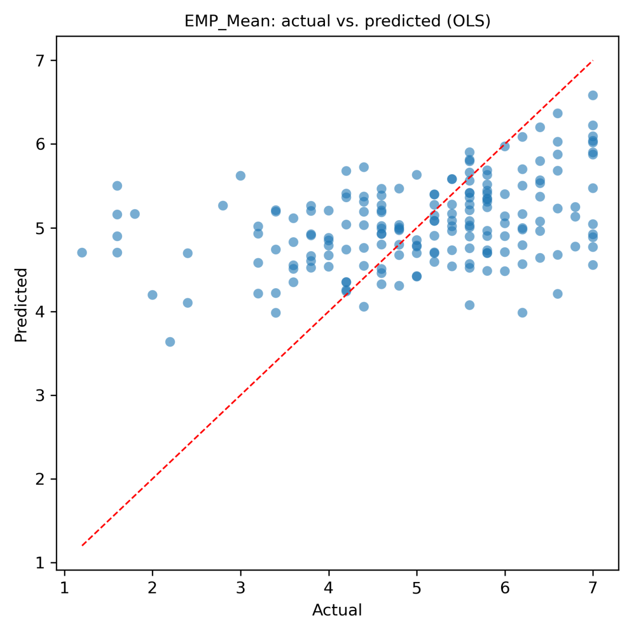



**Figure 7: Observed versus OLS-predicted perceived empathy. Corresponds to Table 5, Model (1).**

Accordingly, the selected survey items used in this study capture key aspects of users' social presence such as perceived conversational quality, empathy and emotional understanding, trust in human-AI interaction.

# C Regression Model Visualization

This section provides supplementary visualizations for the OLS models summarized in Table 5, including Fig 6, Fig 7, Fig 8, Fig 9, Fig 10 and Fig 11.

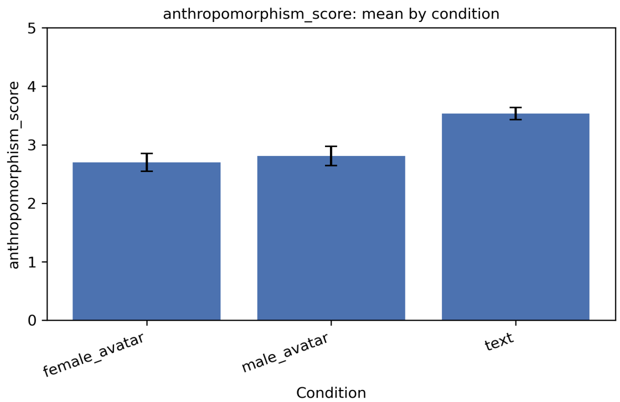



**Figure 10: Mean composite anthropomorphism score by condition. Corresponds to Table 5, Model (3).**

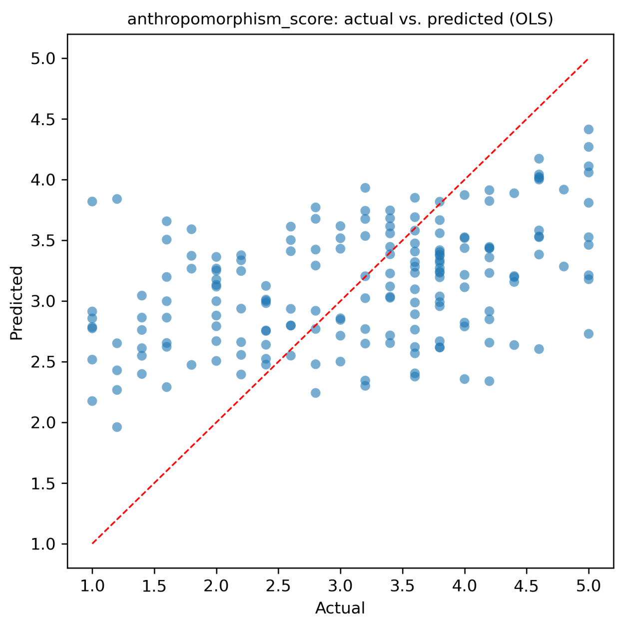



**Figure 11: Observed versus OLS-predicted anthropomorphism. Corresponds to Table 5, Model (3).**

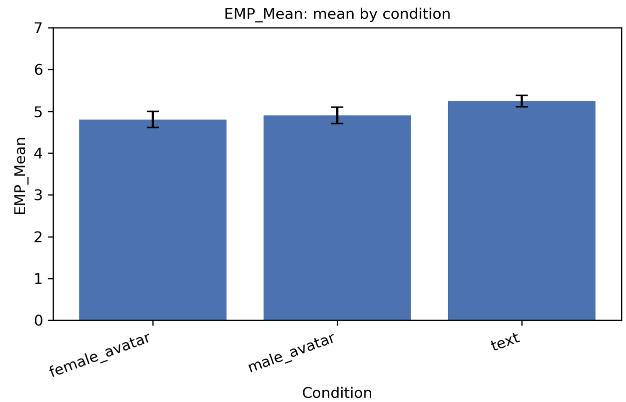



**Figure 8: Mean perceived empathy (EMP composite) by condition. Corresponds to Table 5, Model (2).**

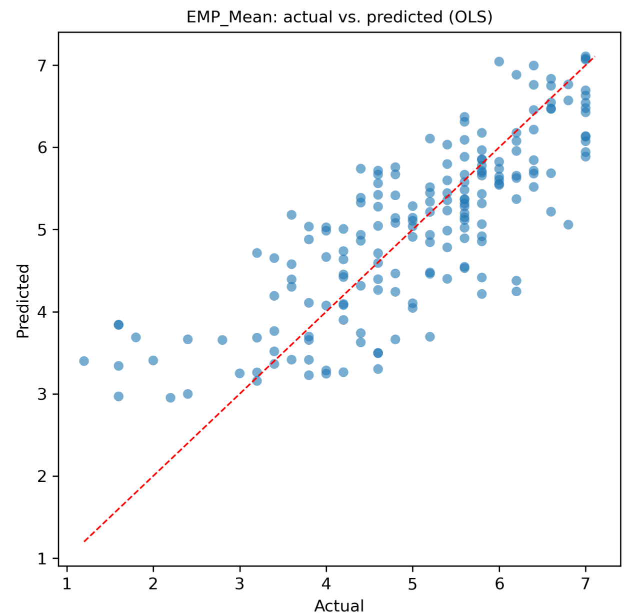



**Figure 9: Observed versus OLS-predicted perceived empathy. Corresponds to Table 5, Model (2).**