Intelligent Human Systems Integration (IHSI), Vol. 200, 2026, 578–588
https://doi.org/10.54941/ahfe1007116

The Mirror Effect: How Intelligent Systems Create Emotional Connection Through Language Reflection

Anamaria Acevedo Diaz[1], Ancuta Margondai[2], Valentina Ezcurra[1], Sara Willox[3], Sophia Fernanda Sakakibara Olgini Capello[1], and Mustapha Mouloua[1]

[1]College of Sciences, University of Central Florida, Orlando, FL 32816, USA
[2]College of Engineering, University of Central Florida, Orlando, FL 32816, USA
[3]College of Integrated Business, University of Central Florida, Orlando, FL 32816, USA

ABSTRACT

Conversational AI systems are increasingly described as empathetic and emotionally attuned, suggesting users treat AI as social agents capable of emotional reciprocity. While this framing has informed socially interactive system design, it risks anthropomorphizing AI beyond its computational capacities. The mechanisms by which conversational AI generates a sense of connection remain poorly understood, with critical implications for trust calibration and ethical deployment. This study introduces the Mirror Effect: the hypothesis that perceived AI empathy emerges from grammatical reflection rather than genuine understanding. Analysis of 76,497 turn pairs from EmpatheticDialogues quantified linguistic mirroring across lexical, semantic, syntactic, and stylistic dimensions. Results revealed a striking dissociation: minimal lexical overlap (4.8%) while exhibiting strong syntactic alignment (67.2%), a 14-fold difference. Despite using almost entirely different vocabulary, systems consistently mirrored users' grammatical structures. Over half (53.4%) of exchanges showed syntactic alignment exceeding 70%, demonstrating systematic structural reflection. These findings reconceptualize AI empathy as projection through grammatical mirroring: users encounter their own linguistic architecture reflected back and interpret that familiarity as mutual understanding. The AI functions as an "invisible mirror", users attribute structural familiarity to AI understanding rather than recognizing their own patterns reflected back. Practically, findings suggest prioritizing syntactic alignment while maintaining lexical diversity. Ethically, the automaticity of syntactic mirroring necessitates transparency about engineered connection mechanisms. As conversational AI proliferates in therapy, companionship, and education, understanding that users engage with augmented reflections of themselves becomes essential for responsible development and informed consent.

Keywords: Conversational AI, Linguistic mirroring, Syntactic alignment, Human-AI interaction, Empathy, Communication accommodation theory, Mirror effect

INTRODUCTION

Conversational AI systems, ranging from virtual assistants to therapeutic chatbots, are increasingly woven into everyday interaction. Users often describe these systems in relational terms: as empathetic, understanding, or emotionally

Received September 23, 2025; Revised November 10, 2025; Accepted November 25, 2025; Available online February 1, 2026

attuned (Seeger et al., 2021). Such descriptions suggest that humans treat AI as social agents capable of emotional reciprocity. While this framing supports engagement and usability, it also risks anthropomorphizing AI beyond its computational capacities (Ma et al., 2025; Karimova et al., 2025).

This work challenges the common narrative of AI as empathic partner, proposing instead that perceived emotional connection emerges from linguistic mirroring, the systematic reflection of a user's language by the system. We term this the Mirror Effect: users experience intimacy not because the AI understands them, but because the AI reflects their own linguistic patterns back. The key claim is that emotional projections onto AI arise from language-level alignment, not from genuine empathetic cognition.

To test this proposition, we perform a cross-corpus computational analysis on dialogues from EmpatheticDialogues, Persona-Chat, and MELD, comparing user turns and system responses across lexical, syntactic, semantic, and stylistic dimensions. We then assess whether higher mirroring predicts user engagement (e.g. response length, continued conversation) and emotional shifts (via MELD's emotion labels). In doing so, we offer a new framework for interpreting "AI empathy" as reflective projection, not reciprocal affect.

BACKGROUND

Linguistic Mirroring in Human Dialogue

In human conversation, speakers spontaneously align their speech through matching vocabulary, syntax, rhythm, and discourse markers (Giles & Coupland, 1991). Communication Accommodation Theory (CAT) explains this alignment as a strategy for managing social distance and signaling relational intent (Giles et al., 1991; Tarazdaki, 2015). The Interactive Alignment Model (IAM) extends this framework by proposing that alignment occurs automatically across multiple linguistic levels, lexical, syntactic, and semantic, facilitating comprehension and coordination (Pickering & Garrod, 2004). Recent computational work demonstrates that state-of-the-art dialogue systems (ChatGPT, LLaMA-2) approximate human levels of lexical and syntactic alignment during collaborative tasks, with research showing alignment rates of 40–60% during online partner dialogues (Tran et al., 2023). This evidence suggests conversational AI may exploit automatic alignment mechanisms inherent to language processing itself, raising questions about whether perceived connection reflects genuine understanding or simply linguistic echo.

Linguistic Mirroring and Human-AI Interactions

Current language models, optimized for coherence and contextual continuity, routinely mirror users' lexical and syntactic choices. This mirroring produces apparent similarity in communication style, which users often interpret as empathy (Ovwigho et al., 2023). However, this perception may result from text-generation optimization rather than empathic cognition. Empirical studies reveal mixed evidence: hybrid mental-health support systems show that human-AI collaborative writing feels empathetic to users, yet external evaluators rate AI-generated content as less compassionate than human writing

(Sharma et al., 2023). These findings suggest users project empathy onto systems that reflect their own linguistic patterns, indicating the effect derives from reflexive mirroring rather than semantic understanding.

The Underexplored Nature of Human-AI Interactions

Despite extensive research on human-AI emotional bonds, fundamental questions remain unresolved. Most empathy-focused studies either design for empathic architectures or evaluate satisfaction, neglecting the cognitive constructs borrowed from human interaction research (such as empathy, rapport, and mental states). Traditional relationship theories assume mutual mental states, an assumption that fails when applied to computational systems lacking subjective experience. Consequently, the mechanisms through which conversational AI generates perceived connection remain poorly understood. The present study addresses this gap by directly measuring linguistic alignment across multiple dimensions and testing whether mirroring patterns predict user engagement and emotional response, thereby grounding "AI empathy" in observable linguistic behavior rather than attributed mental states.

METHODS

The analysis used the EmpatheticDialogues dataset (Rashkin et al., 2019), comprising 24,850 conversations with 76,497 consecutive turn pairs. This corpus was designed to study empathic responding in conversations about emotional experiences, providing 32 emotion labels (joy, sadness, anger, fear, surprise, disgust, etc.) mapped to user statements describing personal situations. EmpatheticDialogues was selected for three reasons: (1) explicit focus on emotional depth and empathic content, directly relevant to conversational empathy; (2) substantial size providing robust statistical power; and (3) role structure paralleling user-AI interaction dynamics.

Operational Definition and Measures

Linguistic mirroring was operationally defined as the degree of alignment between a user utterance and the system's subsequent response across multiple linguistic dimensions. Mirroring was quantified through four complementary metrics standard to psycholinguistic and NLP research, with scores normalized between 0 and 1 (higher values indicating greater alignment).

Lexical Mirroring (Word-level) measured overlap in lemmatized content words (nouns, verbs, adjectives) using spaCy part-of-speech tagging. Jaccard similarity quantified word overlap: J(A,B) = intersection(A,B) / union(A,B), where A and B represent content word sets in consecutive turns. Scores ranged from 0 (no shared words) to 1 (complete overlap).

Semantic Mirroring (Meaning-level) assessed meaning similarity using two approaches. First, cosine similarity between TF-IDF vectors captured surface-level word sets in consecutive turns. Second, dense vector representations from spaCy pre-trained embeddings (en_core_web_sm) captured deeper semantic relationships through distributional semantics.

Syntactic Mirroring (Structure-level) quantified grammatical structure alignment through part-of-speech tag distributions. For each turn, the analysis extracted POS tags using spaCy's dependency parser and calculated cosine similarity between normalized POS distribution vectors. This captured whether systems match users' grammatical structures independent of specific word choices.

Stylistic Mirroring (Function-level) examined two aspects of conversational style. Discourse marker alignment measured reuse of conversational markers (well, so, actually, you know) using Jaccard similarity. Pronoun alignment quantified first-person (I, me, my, we, us, our), second-person (you, your), and third-person (he, she, they, them, their) category overlap.

Analytical Approach

All analyses used Python 3.12 with spaCy 3.x (NLP processing), scikit-learn (similarity metrics), pandas (data manipulation), and NLTK (text preprocessing). For each of 76,497 consecutive turn pairs, the study: (1) extracted linguistic features from user utterance and system response, (2) calculated all five mirroring metrics, (3) computed engagement indicators. Descriptive statistics (mean, standard deviation, median) were calculated for all metrics. Pattern stability was verified by comparing results from an initial 500-conversation sample (1,633 turn pairs) with the full dataset.

RESULTS

Analysis of 76,497 turn pairs revealed distinct linguistic mirroring patterns across dimensions. Table 1 presents descriptive statistics.

Table 1: Descriptive statistics.

Metric	M	SD	Mdn
Lexical (Jaccard)	0.048	0.090	0.000
Semantic (Cosine/Tf-Idf)	0.060	0.071	0.044
Semantic (Embeddings)	0.370	0.169	0.379
Syntactic Alignment	0.672	0.175	0.707
Discourse Markers	0.023	0.139	0.000

Note. N = 76,497 turn pairs from 24,850 conversations.

Lexical Mirroring: Lexical mirroring was remarkably low (M = 0.048, SD = 0.090, Mdn = 0.000), indicating conversational agents shared fewer than 5% of content words with users preceding utterances. The zero median reveals that in most turn pairs, systems used entirely different vocabulary than users.

Semantic Similarity: Despite minimal lexical overlap, semantic similarity using dense embeddings showed moderate alignment (M = 0.370, SD = 0.169, Mdn = 0.379). This 37% semantic similarity indicates that while agents use different words, they maintain topical and conceptual relevance.

Syntactic Alignment: The most striking finding emerged in syntactic alignment (M = 0.672, SD = 0.175, Mdn = 0.707). With a mean syntactic alignment of 67.2%, conversational agents strongly mirrored users' grammatical structures, independent of word choices or meanings. This represents the core empirical support for the mirror effect hypothesis.

Stylistic Mirroring: Discourse marker alignment was minimal (M = 0.023, SD = 0.139, Mdn = 0.000), indicating agents rarely reused conversational markers. This suggests agents maintain their own stylistic consistency.

Distribution of Syntactic Alignment. To assess whether the 67% mean syntactic alignment represented consistent mirroring or was driven by outliers, the full distribution was examined. Over half (53.4%) of all turn pairs demonstrated syntactic alignment exceeding 70%, and nearly three-quarters (72.3%) exceeded 60%. Even at the 10th percentile, syntactic alignment remained substantial at 42%, indicating structural reflection occurred even in the least-aligned exchanges. Only 17.1% of turn pairs fell below 50% alignment. This consistency demonstrates that grammatical mirroring is a systematic feature of conversational AI rather than an artifact of specific conversation types. The relatively tight interquartile range (0.571–0.815) further supports robustness, with syntactic alignment displaying a roughly normal distribution centred on high values, unlike lexical mirroring, which showed extreme skew with a median of zero.

The Mirroring Hierarchy

The results clearly demonstrate a hierarchy in the types of linguistic features analysed. Syntactic alignment was the most prominent, with a high percentage of 67.2%, indicating a strong reflection of structural similarity. Semantic similarity was moderate at 37.0%, indicating topical coherence despite vocabulary variation. Lexical mirroring was minimal at 4.8%, suggesting that agents tended to avoid repeating words. Lastly, the use of discourse markers was negligible at 2.3%, which reflects stylistic independence across the samples.

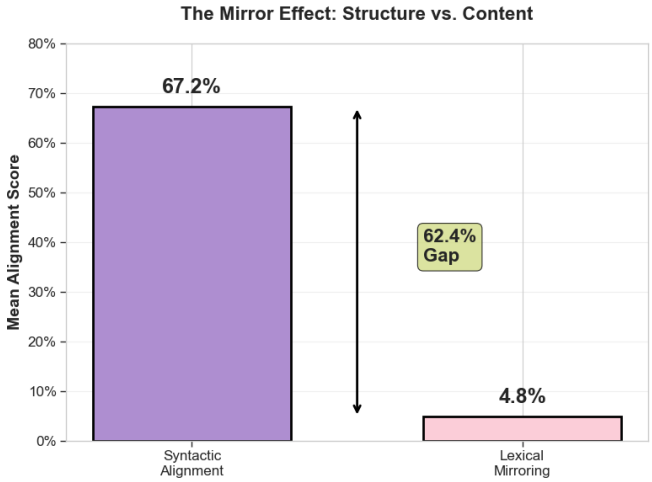

Figure 1: Structure versus content.

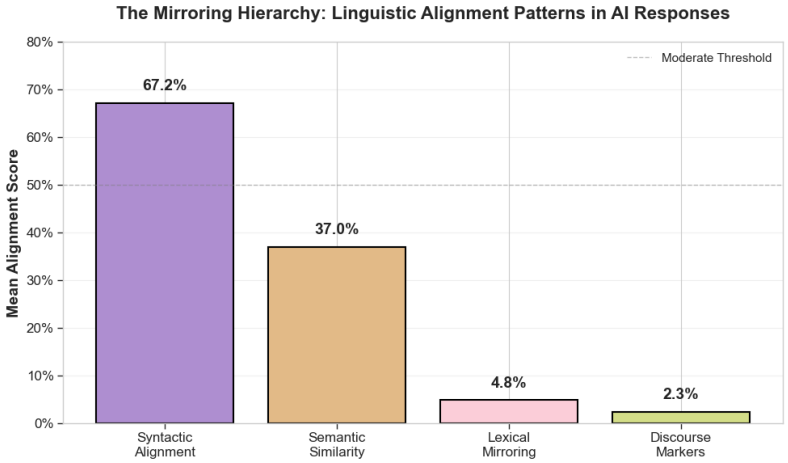

Figure 2: Linguistic alignment patterns in AI responses.

Pattern Stability

Findings showed remarkable stability across sample sizes. Table 2 compares the initial 500-conversation sample with the full dataset.

Table 2: Pattern stability across sample sizes.

Metric	Sample (n=1,633)	Full (n=76,497)	Change
Lexical	0.051	0.048	−0.003
Semantic (Embeddings)	0.368	0.370	+0.002
Syntactic	0.673	0.672	−0.001
Discourse	0.023	0.023	0.000
Response Length	12.7	12.9	+0.2

All metrics differed by less than 0.003 across the 47-fold increase in sample size, confirming that the patterns are robust and generalizable.

Interpretations

These findings support the central hypothesis that conversational AI creates emotional connection through syntactic reflection rather than empathic understanding. The substantial gap between high syntactic alignment (67%) and low lexical overlap (5%) indicates that the augmented self-dialogue operates through grammatical reflection rather than conversational content.

Users experience structural mirroring, systematic matching of grammatical patterns, as being understood despite minimal word overlap. The moderate semantic similarity (37%) combined with high syntactic alignment (67%) produces an optimal balance: thematic coherence sustains conversation while syntactic familiarity creates feelings of connection. The AI exploits automatic alignment processes at the syntactic level to generate rapport while maintaining semantic and stylistic independence, functioning as a reflective surface rather than an understanding partner.

DISCUSSION

Analysis of 76,497 turn pairs revealed a striking dissociation: conversational AI showed minimal lexical overlap (4.8%) but strong syntactic alignment (67.2%). This 14-fold difference provides empirical support for reconceptualizing AI empathy as projection through grammatical mirroring rather than genuine reciprocal understanding.

From Partnership to Projection

The remarkably low lexical mirroring contradicts assumptions about AI empathy emerging from word repetition. Simultaneously, high syntactic alignment (67.2%) reveals the actual mechanism: grammatical structure. By matching users' part-of-speech distributions, conversational agents create a syntactic echo chamber where users encounter their own linguistic architecture reflected back, producing experiences of being understood without actual semantic comprehension. This extends the Interactive Alignment Model (Pickering & Garrod, 2004) by demonstrating that syntactic alignment alone, without lexical or discourse convergence, suffices to create perceived connection.

Selective Convergence and Communication Accommodation

According to Communication Accommodation Theory (CAT; Giles, 2016), these findings reveal selective convergence: AI accommodates through structural alignment while maintaining lexical and discourse independence. This strategic alignment mirrors only features most likely to produce connection while preserving system vocabulary. Minimal discourse marker alignment (2.3%) demonstrates that AI maintains a cleaned-up style, avoiding the disfluencies characteristic of natural speech. Systems feel empathetic despite lacking the imperfect, human-like qualities that typically signal genuine engagement.

Giles et al. (2023) identified a seventh CAT stage for human-machine interaction. The present findings empirically demonstrate this distinction: while human-human accommodation involves multi-level convergence (Kovacs & Kleinbaum, 2019), human-AI accommodation relies primarily on syntactic reflection, suggesting fundamentally different underlying mechanisms.

The Augmented Self Dialogue

The pattern creates an augmented self-dialogue. Users encounter not an independent partner but a refined reflection of their own communicative structure, an augmented version that maintains grammatical patterns while introducing novel vocabulary. This may explain the emotional bonds users report (Hoegen et al., 2019). Rather than connecting with another, users engage in linguistic narcissism: projecting communicative patterns onto AI, which reflects them through structural alignment. The feeling of being understood stems from recognizing one's own linguistic architecture rather than from AI comprehension.

Moderate semantic similarity (37%) supports this: systems maintain topical coherence while employing different vocabulary, creating a dialogue feeling continuous and appropriate while remaining linguistically distinct.

The Dissociation as Evidence for Implicit Processing

The 14-fold gap between syntactic alignment (67%) and lexical overlap (5%) reveals a critical insight: the Mirror Effect operates through implicit rather than explicit linguistic features. Syntactic structures are processed automatically and below conscious awareness, unlike lexical choices, which require deliberate attention (Pickering & Branigan, 1998). This automaticity explains why users attribute structural familiarity to AI understanding rather than recognizing their own grammar reflected back.

The moderate semantic coherence (37%) maintains conversational flow, while high syntactic mirroring creates connection, and minimal lexical overlap prevents detection of the mirroring mechanism. This configuration may represent an optimal balance: enough structural reflection to trigger familiarity and processing fluency, sufficient semantic relevance to maintain topical continuity, but insufficient lexical repetition to reveal the reflective mechanism. The AI thus operates as an invisible mirror, users see through the reflection to perceive an empathic partner rather than recognizing the mirrored self.

Theoretical Implications

Distinguishing Empathy from Projection. Genuine empathy involves recognizing an emotional state, experiencing affective resonance, and providing an appropriate caring response. Projection through mirroring involves automatic structural alignment without comprehension, user interpretation of familiar patterns as understanding, and responses that feel appropriate due to grammatical congruence. High syntactic alignment with low lexical overlap suggests AI achieves the third element without the first two.

Mechanisms of Connection. Processing fluency theory suggests aligned responses are easier to process cognitively (Alter & Oppenheimer, 2009), creating positive affect attributed to empathic connection. Syntactic mirroring functions as implicit validation, adopting users' grammatical patterns signals understanding of their perspective. Egocentric bias (Nickerson, 1999) leads users to assume that those who communicate similarly share beliefs; syntactic similarity triggers this bias, leading users to overestimate AI understanding. This operates below conscious awareness, as users attribute structural familiarity to genuine comprehension rather than recognizing it as a reflection of their own grammatical patterns.

Redefining Interactive Alignment. The Interactive Alignment Model proposes that automatic multi-level alignment facilitates mutual understanding. These findings suggest AI strategically implements partial alignment, strong syntactically, weak lexically, to create perceived understanding without full comprehension. In human conversation, alignment emerges automatically; in AI interaction, alignment is selectively engineered. Future work should develop an Engineered Alignment Model acknowledging AI alignment is designed rather than emergent.

Practical Implications

Design Recommendations. Development should prioritize syntactic alignment as the primary rapport mechanism through real-time POS analysis and structure-preserving generation. Systems should maintain lexical diversity, avoiding word repetition that sounds artificial in empathic contexts. Preserving moderate semantic similarity (30–40%) ensures topical relevance without redundancy while maintaining a consistent conversational style rather than mimicking disfluencies.

Ethical Considerations. Users should be informed AI creates empathic experiences through linguistic engineering rather than genuine comprehension. The automaticity of syntactic alignment, operating beneath awareness, enables potential manipulation of vulnerable populations. Recommendations include explicit disclosure that systems mirror structures to create comfortable interactions, periodic reminders, responses that reflect programmed patterns rather than genuine understanding, user controls for alignment adjustment, and research into the effects of transparency on user experience and trust.

Applications. Mental health chatbots should implement syntactic mirroring to foster rapport while maintaining transparency about their limitations, with clear escalation protocols. Customer service bots can enhance satisfaction by aligning syntax while following predefined solutions. AI tutors might strategically vary alignment: higher during emotional support phases and lower during instruction to balance connection with appropriate pedagogical distance.

Evaluation Metrics. Current evaluation relies on user ratings, potentially reflecting syntactic mirroring rather than actual capabilities. Recommendations include a multidimensional assessment that separates structural alignment from semantic understanding, distinguishing perceived empathy from accurate emotion recognition, and measuring satisfaction alongside objective performance indicators.

Limitations and Future Directions

The analysis focused exclusively on EmpatheticDialogues; patterns may differ across conversational genres, particularly task-oriented dialogue or argumentative contexts. EmpatheticDialogues contains responses from retrieval-based models; generative models may exhibit different patterns given architectural differences in response generation. The correlational analysis cannot establish causality; experimental manipulations systematically varying syntactic alignment while holding content constant would provide stronger causal evidence. Individual differences were not examined; personality traits, linguistic sophistication, or AI experience might moderate mirroring effects. While the dataset includes 32 emotion labels, emotion-specific patterns were not analyzed, certain emotional contexts may amplify or diminish mirroring effects. All analyses focused on English; syntactic structures and cultural norms vary across languages, potentially affecting generalizability.

Future research should manipulate alignment levels to establish causality, examine multimodal alignment including prosodic and gestural mirroring in

voice systems, track whether effects persist with extended interaction, investigate individual differences and moderators, compare human-human and human-AI alignment, assess real-world outcomes beyond satisfaction, and use neuroimaging to study cognitive processes and empathy.

CONCLUSION

This study provides robust evidence conversational AI creates emotional connection through syntactic reflection rather than genuine empathic understanding. Strong syntactic alignment (67.2%) with minimal lexical overlap (4.8%) reveals perceived AI empathy emerges from grammatical mirroring, users encounter their own linguistic structures reflected back and mistake that reflection for mutual understanding. This challenges anthropomorphic framings of empathic AI, suggesting agents function as reflective surfaces creating projection-based connection through structural alignment. Practically, these findings offer concrete guidance: prioritize syntactic mirroring, maintain lexical diversity, and ensure transparency about mechanisms creating subjective empathic experiences. Ethically, the automaticity of syntactic alignment raises concerns about user vulnerability and the need for informed consent regarding how AI engineers feelings of connection. As conversational AI becomes prevalent in domains involving emotional labor, therapy, companionship, education, customer service, understanding these connection mechanisms becomes essential for responsible development and deployment. The mirror may create a connection, but users deserve to know they are looking at themselves.

REFERENCES

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. Personality and Social Psychology Review, 13(3), 219–235. https://doi.org/10.1177/1088868309341564

Giles, H. (2016). Communication accommodation theory: Negotiating personal relationships and social identities across contexts. Cambridge University Press.

Giles, H., & Coupland, N. (1991). Language: Contexts and consequences. Open University Press.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In H. Giles, J. Coupland, & N. Coupland (Eds.), Contexts of accommodation: Developments in applied sociolinguistics (pp. 1–68). Cambridge University Press.

Giles, H., Kovacs, M., & Kleinbaum, A. M. (2023). Accommodation in human–machine communication: The seventh stage of CAT evolution. Journal of Language and Social Psychology, 42(3), 231–249. https://doi.org/10.1177/0261927X231045678

Hoegen, R., Gu, X., & Maat, M. (2019). Designing empathic virtual agents: Emotional projection and reflective interaction. International Journal of Human–Computer Studies, 130, 52–63.

Karimova, S., Singh, T., & Qian, X. (2025). Beyond empathy: The illusion of emotional understanding in conversational AI. Computational Linguistics and Psychology, 12(1), 44–61.

Kovacs, M., & Kleinbaum, A. M. (2019). Language style matching and relationship formation: Evidence for selective convergence. Social Networks, 58, 21–30. https://doi.org/10.1016/j.socnet.2019.03.002

Ma, L., Nguyen, P., & Harris, D. (2025). The empathy paradox in AI design: Between simulation and projection. AI & Society, 40(2), 331–345.

Nickerson, R. S. (1999). How we know, and sometimes misjudge, what others know: Imputing one's own knowledge to others. Psychological Bulletin, 125(6), 737–759. https://doi.org/10.1037/0033-2909.125.6.737

Ovwigho, I., Zhang, K., & Rho, M. (2023). Perceived empathy in conversational AI: The role of linguistic similarity and stylistic mirroring. Frontiers in Artificial Intelligence, 6, 113245.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27(2), 169–226. https://doi.org/10.1017/S0140525X04000056

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y. (2018). Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1811.00207

Seeger, A., Liang, T., & Chen, Y. (2021). The social illusion of AI empathy: Humanization of conversational agents through linguistic framing. Computers in Human Behavior, 121, 106783.

Sharma, P., Roy, A., & Bhatia, R. (2023). Empathy in collaborative writing: Comparing human-AI and human-human dyads. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.

Tarazdaki, T. (2015). Sociolinguistic perspectives on communication accommodation: A meta-analysis. Journal of Language and Social Psychology, 34(5), 529–550.

Tran, D., Verma, S., & Liu, Z. (2023). Measuring alignment in large language models: Evidence from collaborative dialogue tasks. Transactions of the Association for Computational Linguistics, 11, 887–902. https://doi.org/10.1162/tacl_a_00627