Social Sciences & Humanities Open 13 (2026) 102702

Contents lists available at ScienceDirect

Social Sciences & Humanities Open

journal homepage: www.sciencedirect.com/journal/social-sciences-and-humanities-open

Regular Article

How conversational styles relate to trust in virtual humans: Evidence from a synthetic dialogue corpus based on Media Equation Theory

Jiyoung Kang *

Division of Communication and Media, Ewha Womans University, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Virtual humans
Conversational style
Trust formation
Media equation theory
Warmth–competence framework
Human–AI interaction
Contextual effects
Synthetic dialogue corpus

ABSTRACT

This study investigates how conversational styles in virtual humans shape social perceptions, trust, and downstream attitudinal outcomes. Drawing on Media Equation Theory and the warmth–competence framework, we developed a controlled synthetic dialogue corpus of 9720 dialogues generated across two large language model families and three interactional contexts (news, customer service, and education). Four conversational style factors—humor, minor mistakes with self-correction, formality, and friendliness—were systematically manipulated to assess their effects on warmth, competence, and human-likeness.

Structural equation modeling revealed that humor and friendliness significantly enhanced warmth, minor mistakes with self-correction increased human-likeness, and formality improved competence while diminishing warmth. In turn, warmth, competence, and human-likeness each predicted trust, which was associated with increased persuasion and relational orientation. Sequential mediation analyses indicated that style effects were transmitted indirectly via social perceptions and trust, while multi-group comparisons identified context-specific pathways: formality was most influential in news contexts, humor and friendliness in customer service, and self-correction in education.

Together, these findings refine theoretical accounts of human–AI interaction by demonstrating that distinct stylistic cues selectively activate different dimensions of social cognition, and that trust formation depends on balanced, contextually congruent style configurations. Methodologically, the study demonstrates the utility of a synthetic dialogue corpus as a scalable and reproducible approach for theory-driven trust research. Practically, the results provide design-relevant insights for virtual news anchors, service chatbots, and educational tutors that must balance professionalism, warmth, and authenticity, while highlighting the need for complementary validation with human users.

1. Introduction

Virtual humans are increasingly deployed across socially consequential domains such as news delivery, customer service, and education, where they function as anchors, assistants, and tutors. In these settings, effectiveness is not determined solely by visual realism or technical sophistication. Rather, how a virtual human communicates—its linguistic style, including the use of humor, politeness, friendliness, or self-correction—plays a critical role in shaping how users interpret its intentions, competence, and credibility. Foundational theories such as the Media Equation (Reeves & Nass, 1996) and the Computers as Social Actors (CASA) paradigm (Nass & Moon, 2000) suggest that people apply social heuristics to mediated agents, inferring interpersonal qualities from language-based cues. Complementing this view, the warmth–competence framework explains how perceptions of benevolence and ability jointly structure social evaluation, with downstream implications for trust and influence (Fiske et al., 2007).

Despite substantial progress, several limitations remain in the literature on conversational agents and virtual humans. First, much prior work examines stylistic features in isolation. Studies show, for instance, that humor can enhance likeability and engagement (Ceha et al., 2021), while politeness strategies improve cooperation and acceptability in human–agent interaction (Kumar et al., 2022). However, real-world interactions rarely involve single cues in isolation. Linguistic styles typically co-occur, creating potential complementarities and trade-offs—particularly between warmth-related signals (e.g., friendliness, humor) and competence-related signals (e.g., formality, error-free delivery). Empirical research has only begun to address how multiple

* Division of Communication and Media Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul, 03760, Republic of Korea.
E-mail address: kangjiyoung@ewha.ac.kr.

https://doi.org/10.1016/j.ssaho.2026.102702
Received 17 September 2025; Received in revised form 9 February 2026; Accepted 17 March 2026
Available online 23 March 2026
2590-2911/© 2026 The Author. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

stylistic cues jointly shape social perception.

Second, although theoretical accounts imply that conversational style influences behavioral outcomes through intermediate social perceptions, empirical studies seldom model these processes as explicit sequential pathways. For example, research on the "machine heuristic" demonstrates that people may sometimes trust computational agents over humans (Sundar & Kim, 2019), yet this work does not specify how distinct linguistic cues are first translated into perceptions of warmth, competence, or human-likeness before affecting trust. As a result, the internal mechanism linking style to trust and subsequent outcomes remains under-specified.

Third, contextual variation has received limited systematic attention. Communication norms and expectations differ markedly across domains. News contexts emphasize neutrality, formality, and error aversion; customer service interactions often reward friendliness and tension-reducing humor; and educational settings require a balance between approachability and perceived expertise. Prior research has examined virtual humans in each of these domains independently, but direct cross-context comparisons remain rare. This gap is theoretically consequential because context shapes what linguistic behaviors are interpreted as appropriate, trustworthy, or persuasive, thereby constraining the effectiveness of stylistic strategies.

Fourth, existing studies largely assume linear effects of conversational cues. However, research on anthropomorphism and social signaling suggests that stylistic intensity may follow nonlinear patterns, in which moderate cues enhance trust and credibility while excessive cues undermine them (Waytz et al., 2014). Humor, self-corrections, or informality may therefore exhibit diminishing returns or even backfire effects, a possibility that has been insufficiently explored in conversational agent research.

Finally, methodological constraints limit cumulative progress. Human-participant experiments are resource-intensive, difficult to scale, and challenging to reproduce when multiple stylistic dimensions and contextual conditions are manipulated simultaneously. Recent advances in large language models (LLMs) have enabled the generation of controlled synthetic dialogue corpora, offering scalability and experimental precision. Importantly, such synthetic approaches are not intended to replace human-subject research, but rather to function as theory-testing and hypothesis stress-testing tools that allow systematic exploration of complex design spaces under tightly controlled conditions. Their validity, however, depends on transparent design decisions and rigorous validation procedures. Recent work has increasingly positioned LLM-based synthetic corpora as tools for controlled mechanism testing and hypothesis exploration in human–AI interaction research, particularly when direct human experimentation is costly or ethically constrained.

Taken together, these limitations indicate that conversational style in virtual humans should be understood not as isolated linguistic cues but as a configurational phenomenon, in which multiple stylistic signals jointly shape social interpretation within specific contextual constraints.

This configurational perspective directly motivates the present study's focus on patterned combinations of conversational cues and their indirect effects on trust-related outcomes across domains.

Accordingly, this study pursues two interrelated objectives.

First, at a theoretical level, it refines Media Equation theory and the warmth–competence framework by examining how configurations of conversational style cues shape perceptions of warmth, competence, and human-likeness, and how these perceptions sequentially translate into trust.

Second, at a methodological level, it demonstrates the utility of a rigorously validated synthetic dialogue corpus as a transparent and scalable tool for theory-driven mechanism testing in virtual human research.

To address these gaps, the present study adopts a controlled synthetic dialogue corpus approach to examine how conversational styles shape trust-related outcomes in virtual humans across contexts. We generate 9720 dialogues in three socially salient domains—news, customer service, and education—using two LLM families, while systematically manipulating four linguistic style dimensions: humor, minor mistakes with self-correction, formality, and friendliness. We test a sequential mediation model in which stylistic cues influence perceptions of warmth, competence, and human-likeness, which in turn shape trust and subsequently predict persuasion and relational orientation.

By systematically manipulating multiple conversational style dimensions across distinct interaction contexts, this study advances prior work that has predominantly focused on single cues or single domains, enabling a configurational and interaction-level understanding of trust in virtual humans.

By integrating Media Equation theory with the warmth–competence framework and situating synthetic corpus methods as a transparent tool for mechanism testing, this study aims to clarify how conversational style operates within specific contextual boundaries, while acknowledging that the observed effects represent theoretically grounded proxies pending validation with human participants. Beyond its theoretical contributions, the study makes a methodological contribution by implementing a multi-layered quality control and validation pipeline and a large-scale, cross-context design that together enable robust, configurational, and interaction-level insights into trust formation in virtual humans.

2. Literature review

2.1. Virtual humans and conversational style

Virtual humans have developed from the tradition of embodied conversational agents (ECAs), which emphasized multimodal cues such as gaze, gesture, and prosody to reproduce naturalistic interaction (Cassell et al., 2000). While early ECA research treated language as secondary to embodiment, recent advances in large language models (LLMs), together with alignment and safety mechanisms, have shifted scholarly attention toward language-centered social signaling. Contemporary research increasingly conceptualizes LLM-driven virtual humans not merely as interfaces, but as socially interpretable agents whose linguistic behavior shapes user perception, acceptance, and trust (Araujo & Bol, 2024; Gerlich, 2024).

These developments have accelerated the deployment of virtual humans in socially sensitive domains such as news, customer service, and education, where conversational style functions as a normative constraint rather than a purely aesthetic choice. Recent studies suggest that trust and legitimacy depend not only on informational accuracy but also on whether linguistic behavior aligns with domain-specific expectations and ethical standards (Gnewuch et al., 2017; Gerlich, 2024). Extending this line of research, recent evidence shows that politeness and linguistic sensitivity are particularly consequential in emotionally sensitive interactions, such as mental health support, where they strongly shape user experience and acceptance of conversational agents (Bowman et al., 2024).

Against this backdrop, prior research identifies four stylistic dimensions as especially relevant for trust-related evaluation: humor, formality, self-correction, and friendliness. Although each cue has been shown to influence engagement, credibility, or relational orientation, their effects are highly contingent on intensity, combination, and contextual fit, and may even backfire when misaligned with institutional norms or interactional expectations (Ceha et al., 2021; Dubiel et al., 2022; Kumar et al., 2022; Araujo & Bol, 2024).

2.2. Trust formation in Human–AI interaction

People routinely treat computers and media interfaces as social actors, applying interpersonal norms to machines on the basis of linguistic and nonverbal cues. This phenomenon, explained by the Media Equation and CASA paradigms (Nass & Moon, 2000; Reeves & Nass, 1996),

suggests that conversational cues provide the foundation for social judgments. Within this perspective, trust in human–AI interaction is not assumed to be inherent or universal, but emerges from users' interpretations of socially meaningful signals. Trust in such systems can be conceptualized through the dimensions of warmth and competence (Fiske et al., 2007), which offer a parsimonious yet flexible framework for capturing perceived benevolence and ability without presuming full human equivalence. Adding human-likeness or perceived authenticity further strengthens explanatory power, as anthropomorphic cues have been shown to increase trust in autonomous and nonhuman systems, particularly when they support intelligibility and transparency rather than mere surface mimicry (Waytz et al., 2014; Lee & Nass, 2004).

Conversational style functions as a primary signaling mechanism through which these dimensions are inferred. Humor, friendliness, and self-correction can enhance perceived warmth and authenticity, while higher levels of formality and politeness reliably signal competence and professionalism, often at the expense of approachability (Dubiel et al., 2022; Kumar et al., 2022). Importantly, these effects are not uniform across situations: stylistic cues are evaluated relative to contextual expectations and institutional norms, such that the same cue may calibrate trust positively in one domain while undermining it in another. This underscores the need to examine trust not as a direct outcome of style, but as a perception-mediated process.

Moreover, trust in human–AI interaction is increasingly understood as dynamic rather than static. Repeated exposure and personalization can transform initial cue-based judgments into more relational forms of trust, shaped by consistency, predictability, and perceived alignment over time (Bickmore & Picard, 2005; Araujo & Bol, 2024). Hoff and Bashir's (2015) synthesis further emphasizes that trust calibration depends on both situational cues and design choices, highlighting the risk of miscalibration when stylistic signals overstate an agent's reliability or social intent. Recent discussions extend this concern to ethical dimensions, noting that overly optimized warmth or friendliness may blur the boundary between appropriate trust calibration and undue persuasion.

Taken together, prior work positions conversational style as a central—but context-bounded—mechanism of trust formation, operating through users' interpretations of warmth, competence, and human-likeness. At the same time, much of the existing evidence relies on human-subject studies or conceptual models, leaving open questions about how these mechanisms interact under tightly controlled manipulations of multiple stylistic cues. In this study, trust is therefore treated as a mediating construct within a theoretically specified model, and its measurement should be interpreted as a proxy outcome derived from synthetic interactions rather than as validated human behavioral trust. Building on this conceptualization, we next examine persuasion and relational orientation as downstream outcomes of trust-related perceptions.

2.3. Persuasion and relational orientation in AI communication

Persuasion in human–AI communication has traditionally been explained through dual-process models of information processing. According to the Elaboration Likelihood Model (ELM; Petty & Cacioppo, 1986), highly motivated individuals process messages via the central route by evaluating argument quality, whereas less motivated individuals rely on peripheral cues such as source credibility or likability. Similarly, the Heuristic–Systematic Model (HSM; Chaiken, 1980; Chaiken et al., 1989) proposes that heuristic and systematic processing can operate in parallel, depending on cognitive resources and situational demands. Within contemporary AI-mediated contexts, conversational agents often function as heuristic sources, particularly when users have limited motivation or expertise, making stylistic cues especially influential in shaping judgments.

Within these frameworks, conversational style cues operate as peripheral signals that bias evaluation. Humor and friendliness enhance perceived warmth and likability, while formality and linguistic precision reinforce competence and professionalism. Crucially, the effectiveness of these cues depends on contextual appropriateness rather than their mere presence, as mismatches between style and situational norms can reduce persuasive impact or even trigger resistance. The influence of source credibility—often inferred from warmth and competence cues—has been consistently documented across persuasion research (Pornpitakpan, 2004). Recent human–AI studies further suggest that credibility judgments toward artificial agents are highly sensitive to linguistic tone and error management, particularly in institutional or authority-laden settings, where credibility expectations are tightly regulated.

Trust functions as a pivotal linking mechanism between perception and persuasion. A meta-analysis in human–robot interaction demonstrates that trust robustly predicts reliance and compliance behaviors (Hancock et al., 2011). In relational agent research, Bickmore and Picard (2005) show that relational orientation—defined by intentions for continued interaction, emotional attachment, and social expectations—emerges through iterative trust-building processes. More recent work emphasizes that such relational outcomes should be interpreted as attitudinal or intentional orientations rather than stable interpersonal bonds, especially in short-term or text-based interactions (Araujo & Bol, 2024).

Taken together, persuasion and relational orientation in human–AI communication are best understood as downstream outcomes of a perception-mediated trust process rather than as direct effects of stylistic cues. Conversational style shapes perceptions of warmth, competence, and human-likeness, which calibrate trust and subsequently influence persuasive acceptance and relational intent. In the present study, persuasion and relational orientation are therefore treated as proxy outcomes derived from synthetic interactions, capturing theoretically grounded tendencies rather than validated human behavioral change. This framing allows systematic testing of hypothesized mechanisms while acknowledging the need for future validation with human participants.

2.4. Context as a moderator

The effectiveness of conversational style is not universal but systematically moderated by contextual conditions. From the perspective of dual-process models such as ELM and HSM, task importance, social norms, and perceived risk shape whether users rely on central evaluation or peripheral cues when interacting with AI agents (Chaiken, 1980; Petty & Cacioppo, 1986). Accordingly, the same stylistic cue may be interpreted as credible, inappropriate, or even manipulative depending on the situational frame in which it appears.

In news contexts, communication is governed by strong institutional norms emphasizing accuracy, neutrality, and professionalism. Here, linguistic formality and error minimization are likely to function as trust-preserving cues by reinforcing perceived competence and legitimacy, while excessive friendliness or humor may violate expectations of objectivity. By contrast, in customer service interactions—often characterized by heightened uncertainty, frustration, or emotional load—warmth-oriented cues such as friendliness and context-appropriate humor can reduce tension and facilitate trust repair, improving satisfaction and compliance (Gnewuch et al., 2018). In educational settings, users frequently tolerate or even value expressions of uncertainty, transparency, and self-correction, as these cues align with pedagogical norms that frame learning as an iterative and fallible process, thereby supporting trust calibration rather than undermining authority (Dubiel et al., 2022).

Media process theories further reinforce the importance of contextual moderation. The MAIN model (Sundar, 2008) posits that interface cues activate heuristics related to agency, credibility, and control, but the strength and direction of these heuristics depend on users' goals and expectations within a given domain. Schema congruity theory similarly

suggests that moderate incongruity between expected and observed agent behavior maximizes attention and engagement, producing nonlinear or inverted-U effects (Meyers-Levy & Tybout, 1989). Applied to virtual humans, this implies that moderate deviations from stereotypical role expectations (e.g., limited warmth in news, mild humor in customer service) may enhance engagement, whereas excessive deviation risks credibility loss.

Evidence from adjacent domains, such as autonomous vehicles and social robots, further supports this view. Anthropomorphic cues have been shown to enhance trust when they align with situational demands, but to undermine confidence when they overstate social or moral competence relative to system capabilities (Waytz et al., 2014). Recent discussions extend this concern to ethical considerations, cautioning that misaligned human-likeness may induce inappropriate trust or reliance, particularly in high-stakes or authority-laden contexts (Gerlich, 2024).

Taken together, these perspectives suggest that conversational style effects should be interpreted as context-bounded rather than universal. This justifies the present study's focus on three domains—news, customer service, and education—that differ systematically in institutional norms, user expectations, and risk profiles. By examining these contexts in parallel, the study treats context not as a background variable but as a theoretically motivated moderator that constrains how stylistic cues are translated into social perceptions and trust.

2.5. Summary of research gaps and hypotheses development

Synthesizing the literature reveals four key limitations. First, most studies focus on single stylistic features, thereby overlooking interaction effects and trade-offs among co-occurring cues. Although some studies have examined isolated mediation processes (e.g., warmth → trust), few have articulated integrated sequential mediation models that simultaneously link multiple social perceptions—warmth, competence, and human-likeness—to trust and downstream outcomes such as persuasion and relational orientation.

Second, contextual moderation has been insufficiently theorized and empirically tested, despite strong theoretical indications that stylistic cues are interpreted relative to domain-specific norms and expectations. In particular, direct cross-context comparisons that treat context as a boundary condition rather than a background variable remain rare.

Third, nonlinear effects of conversational style—such as inverted-U patterns predicted by schema congruity and anthropomorphism research—have received limited empirical attention in human–AI interaction, even though excessive warmth, humor, or human-likeness may undermine credibility and trust.

Fourth, existing research relies predominantly on small-scale human experiments, which constrains reproducibility, limits the systematic manipulation of multiple stylistic dimensions, and complicates cross-context comparison. These constraints are further amplified by ethical and logistical challenges in participant recruitment, especially in high-stakes or authority-laden contexts (Hoff & Bashir, 2015).

The present study addresses these gaps through a controlled synthetic dialogue corpus and factorial design. By systematically manipulating humor, formality, minor mistakes with self-correction, and friendliness across news, customer service, and education contexts, the study is designed to stress-test theoretically specified mechanisms rather than to substitute for human behavioral validation. We examine direct, indirect, and moderated effects of conversational style on social perceptions (warmth, competence, human-likeness), trust, persuasion, and relational orientation. Within this framework, trust, persuasion, and relational orientation are treated as theoretically grounded proxy outcomes derived from synthetic interactions.

This approach enables scalable and reproducible testing of complex interaction patterns, including both linear and nonlinear relationships, while explicitly acknowledging that the resulting effect estimates represent model-based tendencies pending confirmation in human-subject research. In doing so, the study advances theoretical clarity regarding style–trust mechanisms and provides a principled foundation for future empirical validation and responsible virtual human design.

3. Research model and hypotheses

3.1. Conceptual framework

Fig. 1 presents the conceptual framework of this study. The model posits that conversational style cues—humor, intentional minor mistakes with self-correction, formality, and friendliness—shape users' perceptions of warmth, competence, and human-likeness/authenticity. These perceptions are conceptualized as proximal, perception-level constructs rather than stable interpersonal evaluations, reflecting users' immediate interpretations of linguistic behavior. These social perceptions, in turn, function as antecedents of trust, which is modeled as a mediating mechanism within the proposed framework rather than as a universal or unconditional outcome. Trust subsequently predicts persuasion and relational orientation, which are treated as attitudinal and intentional outcomes within the confines of the present model.

Context (news, customer service, education) is modeled as a moderator, reflecting that communicative norms, institutional roles, and task priorities differ substantially across domains (Fiske et al., 2007; Nass & Moon, 2000; Reeves & Nass, 1996). Rather than serving as a background variable, context is theorized as a boundary condition that constrains how stylistic cues are translated into social perceptions and trust. This approach acknowledges that the same linguistic behavior may signal credibility, appropriateness, or impropriety depending on domain-specific expectations.

The framework integrates insights from social cognition and persuasion theory. The warmth–competence model explains how people infer intentions and ability from stylistic cues, providing a parsimonious structure for understanding social perception without presuming full human equivalence (Fiske et al., 2007). Dual-process models of persuasion, including the Elaboration Likelihood Model (Petty & Cacioppo, 1986) and the Heuristic–Systematic Model (Chaiken, 1980), suggest that conversational cues can operate as peripheral signals (e.g., humor signaling friendliness) or bias systematic evaluation (e.g., formality reinforcing competence), particularly under conditions of limited motivation, expertise, or cognitive resources.

Repeated and personalized interactions have been shown to transform initial cue-based judgments into more relationally oriented evaluations (Bickmore & Picard, 2005; Araujo & Bol, 2024). However, within the present study's synthetic and text-based design, such relational outcomes are conceptualized as theoretically grounded proxies rather than empirically validated long-term bonds. Finally, contextual theories such as the MAIN model (Sundar, 2008) and schema congruity theory (Meyers-Levy & Tybout, 1989) highlight that the salience, direction, and optimal intensity of stylistic cues are contingent on situational norms, allowing for the possibility of nonlinear or inverted-U effects.

Together, these perspectives motivate a sequential, perception-mediated framework—conversational style → social perceptions → trust → persuasion and relational orientation—in which context moderates key pathways. The proposed model is intended to test theoretically specified mechanisms under controlled conditions, with the understanding that observed relationships reflect model-based tendencies that warrant subsequent validation in human-subject research.

3.2. Hypotheses

3.2.1. Conversational style and social perceptions

Conversational cues serve as immediate signals for social perception. Humor and friendliness are consistently associated with greater warmth and interpersonal appeal, and affiliative humor has been shown to increase user engagement and perceived authenticity in human–agent

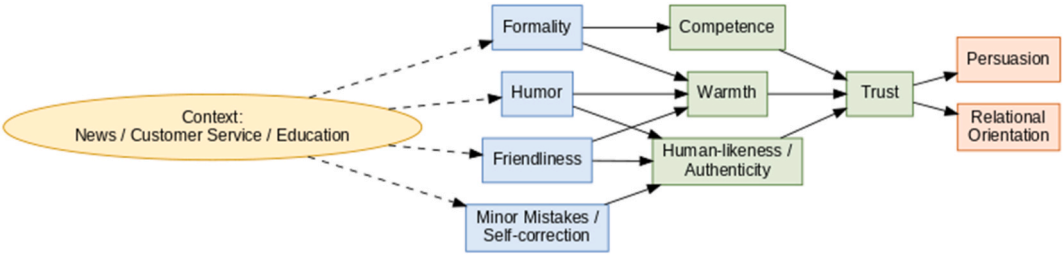


Fig. 1. Conceptual model of the study.

interactions (Ceha et al., 2021; Niculescu & Banchs, 2019). Minor mistakes accompanied by self-correction provide cues of transparency and fallibility, which can increase perceived human-likeness by mimicking everyday conversational repair (Dubiel et al., 2022). By contrast, formality conveys professionalism and competence, aligning with politeness theory and findings in HRI (Kumar et al., 2022), but may attenuate warmth when interpreted as socially distant. rigid or socially distant.

- **H1a.** Humor and friendliness will positively influence perceptions of warmth and human-likeness/authenticity.
- **H1b.** Intentional minor mistakes followed by self-correction will increase perceptions of human-likeness/authenticity.
- **H1c.** Formality will positively influence competence but negatively influence warmth.

3.2.2. Social perceptions and trust

Warmth, competence, and human-likeness are theorized to predict trust in virtual humans. Warmth signals benevolence, competence signals capability, and human-likeness signals authenticity—together anchoring social judgments (Fiske et al., 2007). Persuasion research similarly identifies credibility, derived from these dimensions, as a key determinant of trust (Pornpitakpan, 2004). Research on automation and HRI suggests that anthropomorphic cues can enhance trust calibration in nonhuman systems, contingent on contextual appropriateness (Waytz et al., 2014; Hoff & Bashir, 2015).

- **H2a.** Perceived warmth will positively predict trust.
- **H2b.** Perceived competence will positively predict trust.
- **H2c.** Perceived human-likeness/authenticity will positively predict trust.

3.2.3. Trust and downstream outcomes

Trust functions as a pivotal mediator between perception and outcome. In persuasion contexts, trusted sources increase message acceptance (Petty & Cacioppo, 1986). In human–robot interaction, trust predicts compliance and reliance (Hancock et al., 2011). Trust also underpins relational orientation, defined as willingness to continue interacting and forming bonds (Bickmore & Picard, 2005).

- **H3a.** Trust will positively predict persuasion (message acceptance).
- **H3b.** Trust will positively predict relational orientation (intention for continued interaction).

3.2.4. Sequential mediation

Integrating these pathways suggests a sequential mediation mechanism in which conversational style influences social perceptions, which shape trust and subsequently predict persuasion and relational orientation. Within the present framework, trust is positioned as a model-specific mediating construct linking stylistic cues to downstream attitudinal outcomes.

- **H4.** Conversational style will indirectly influence persuasion and relational orientation through the sequential pathway of social perceptions → trust.

3.2.5. Contextual moderation

The strength of these pathways is expected to vary across domains. In news contexts, where professionalism and accuracy are paramount, formality is expected to exert a stronger influence on competence and trust. In customer service, humor and friendliness are particularly effective in reducing tension and enhancing warmth, thereby strengthening trust (Gnewuch et al., 2018). In educational contexts, transparency and self-correction play a central role in credibility formation and trust calibration (Dubiel et al., 2022).

- **H5a.** In news contexts, the formality → competence → trust pathway will be stronger than in customer service or education contexts.
- **H5b.** In customer service contexts, the humor/friendliness → warmth → trust pathway will be stronger than in news or education contexts.
- **H5c.** In educational contexts, the self-correction → human-likeness/authenticity → trust pathway will be stronger than in other contexts.

3.3. Exploratory research questions

Beyond the formal hypotheses, two exploratory issues warrant investigation. First, stylistic effects may not be strictly linear. Prior research on humor, anthropomorphism, and social signaling suggests inverted-U dynamics, in which moderate levels of cues optimize warmth and trust, whereas excessive use may reduce perceived competence or credibility (Meyers-Levy & Tybout, 1989; Waytz et al., 2014). Such nonlinear patterns are theoretically plausible in human–AI interaction, yet remain empirically underexplored under controlled, multi-style manipulations.

Second, conversational style cues may interact, producing complementary or offsetting effects rather than independent influences. For example, humor may soften the rigidity associated with high formality, while friendliness may amplify the positive effects of humor on perceived warmth. Examining these interactions allows assessment of whether certain stylistic combinations mitigate trade-offs between warmth and competence or instead introduce new forms of perceptual inconsistency.

- **RQ1.** Do humor, formality, and minor mistakes with self-correction exhibit inverted-U relationships with outcomes, indicating optimal effects at moderate levels?
- **RQ2.** Do style–style interactions—specifically humor × formality, humor × mistakes, and friendliness × formality—produce complementary or offsetting effects on perceptions, trust, and outcomes?

4. Methodology

4.1. Research design

Building directly on the theoretical framework and hypotheses developed in the preceding sections, this study adopted a synthetic corpus–based experimental design to test the sequential process whereby conversational style influences social perceptions, which then shape trust and downstream outcomes of persuasion and relational orientation. The use of a fully synthetic corpus addresses methodological constraints in prior research—particularly the reliance on small-scale participant studies that limit reproducibility and cross-context comparison (Hancock et al., 2011; Hoff & Bashir, 2015). By systematically generating and manipulating dialogues across multiple conditions, the design enables large-scale, reproducible, and ethically unobtrusive testing of stylistic effects.

The unit of analysis was a fixed nine-turn dialogue, defined as an alternating interaction consisting of four user turns and five assistant turns, with each turn averaging two to three sentences. The turn sequence was held constant across all experimental conditions to ensure structural equivalence, with conversational style manipulations applied exclusively to assistant turns. User turns were generated from a predefined template and remained identical across conditions, such that any observed variation in social perceptions and outcomes could be attributed solely to differences in assistant language style. Detailed prompt templates, turn protocols, and an exemplar nine-turn dialogue are provided in Appendix A.

A fixed fact skeleton specified the informational content of all dialogues, ensuring semantic equivalence across conditions and allowing only conversational style variables to vary. Four style factors were manipulated: humor, intentional minor mistakes with self-correction, formality, and friendliness. Style manipulations were implemented as pre-specified constraints applied consistently across conditions. This isolation of linguistic style accords with Media Equation Theory, which holds that social judgments can be elicited from linguistic cues alone (Reeves & Nass, 1996), and with the warmth–competence framework, which treats style as a determinant of perceived intentions and ability (Fiske et al., 2007).

To enhance robustness across systems, dialogues were generated using two large language model (LLM) families—GPT-4–class models (OpenAI, 2023) and Claude-3–class models (Anthropic, 2024). These model families were treated as parallel generation sources rather than analytic factors, reducing the likelihood that observed effects reflect idiosyncrasies of a single architecture and facilitating cross-system replication. A class-level specification was adopted to ensure longitudinal reproducibility in light of rapid version updates in commercial LLM APIs.

Dialogue generation was conducted using fixed prompt templates, predefined stylistic constraints, and controlled decoding parameter ranges (temperature = .4–.7; top-p = 0.8–1.0). Random seeds were fixed at the batch level during regeneration to ensure that replacement dialogues preserved the intended stylistic configuration. Default system and safety policies of the respective LLM APIs were retained and not overridden during generation.

A 3 × 3 × 3 × 2 × 3 factorial design was implemented. As summarized in Table 1, humor was manipulated at three levels (none, moderate, high), minor mistakes/self-correction at three levels (none, one occurrence, two occurrences with rationale), formality at three levels (low, medium, high), friendliness at two levels (low, high), and context at three levels (news, customer service, education). This full factorial design yielded 162 experimental cells, with 30 dialogues generated per cell. For each model family, the sample size was N = 4860; combined, N = 9720. Randomization was applied at the cell level, and decoding parameters were sampled within constrained ranges (temperature .4–.7, top-p .8–1.0) to balance diversity and reproducibility. Duplicate outputs were removed using Sentence-BERT cosine similarity (threshold ≥.92).

Table 1
Factorial design and cell counts(per model family).

Factor (levels)	Levels	Notes
Humor (3)	none/moderate/high	Context-matched wit and wordplay intensity control
Minor mistakes & self-correction (3)	none/1 occurrence/2 occurrences + rationale	No conflict with fact skeleton
Formality (3)	low/medium/high	Colloquial → standard → formal/technical register
Friendliness (2)	low/high	Use of name, empathy, self-disclosure frequency
Context (3)	news/customer service/education	Distinct task norms and goals

A consolidated overview of the dialogue generation, quality control, and inclusion pipeline—summarizing key generation parameters, validation thresholds, and exclusion criteria—is provided in Appendix A.0 (Table A0).

Because the corpus consisted entirely of machine-generated text and involved no human participants or personal data, the study qualifies for institutional review board (IRB) exemption. In this way, the methodological design directly responds to the challenges identified in Section 2.5, offering a scalable and reproducible alternative to traditional lab-based experiments.

4.2. Stimulus generation

Stimuli were generated through carefully designed prompts that operationalized each stylistic factor. A prompt skeleton specified the role, task, and context of the dialogue, as well as style rules, lexical constraints, and turn-taking structure. All prompts were fixed ex ante and applied identically across experimental cells, with no adaptive or iterative prompt modification during generation. Each factor was operationalized through explicit linguistic rules to ensure construct validity. For example, humor levels were defined by the frequency and type of wit, while self-correction was implemented as one or two deliberate errors with immediate rectification using pre-specified correction markers.

To further enhance internal validity, all dialogues were presented in text-only format, excluding visual representations of virtual humans. This design choice prevents potential confounds from visual cues such as avatars, facial expressions, or background, thereby ensuring that measured perceptions of warmth, competence, and human-likeness reflect only linguistic manipulation. By holding modality constant, the design isolates language-based social signaling as the sole source of variance. This decision is theoretically supported by Media Equation Theory (Reeves & Nass, 1996) and the warmth–competence framework (Fiske et al., 2007), as well as empirical findings that conversational style alone can elicit social judgments (Niculescu & Banchs, 2019; Dubiel et al., 2022).

The detailed operationalization of each style factor is presented in Table 2. Humor ranged from prohibited (Level 0) to context-appropriate wit (Level 1) and more frequent metaphor/wordplay (Level 2), with an upper bound of one humorous expression per turn to prevent uncontrolled stylistic escalation. Self-correction ranged from none (Level 0) to one immediate correction (Level 1) and two corrections with a brief rationale (Level 2), restricted to minor numerical or lexical deviations that did not alter the fact skeleton. Formality was manipulated across colloquial, standard, and technical registers, with register-specific lexical lists used to enforce consistency. Friendliness was defined by minimal versus frequent use of the addressee's name, empathy expressions, and micro self-disclosures, subject to predefined frequency ceilings per dialogue. These specifications ensured consistency across dialogues and grounded the manipulations in theoretically meaningful constructs.

By structuring the stimuli in this way, the study achieves high internal validity: any differences in outcomes can be attributed directly to

Table 2
Style manipulation levels and prompt guidelines.

Construct	Level	Prompted behaviors (summary)
Humor	0	Humor prohibited; maintain an information-focused tone
	1	Short, context-appropriate wit (≤1 per turn); no disruption to information flow
	2	Use of metaphors and wordplay (≤1 per turn); avoid excessive or off-topic humor
Minor mistakes & self-correction	0	No mistakes or corrections
	1	One minor numerical/lexical mistake → immediate correction ("Correction: …")
	2	Two such mistakes + one-sentence rationale for the correction (emphasizing transparency)
Formality	0	Colloquial style; horizontal address; concise responses
	1	Standard narrative style; polite register; clear sentence structure
	2	Formal/technical register; complete honorific forms; no redundant honorifics
Friendliness	0	Minimal use of addressee's name; restrained emotional expression and self-disclosure
	1	Use of addressee's name; empathy; micro self-disclosures within frequency limits

the manipulated style factors rather than uncontrolled linguistic or visual variation. Moreover, the explicit specification of style rules, lexical constraints, and turn-level frequency bounds—along with dictionaries of prohibited and recommended expressions (see Appendix) supports transparency and replicability, enabling future researchers to reproduce or extend the design under comparable conditions.

4.3. Quality control procedures

Given the scale and synthetic nature of the corpus, a rigorous quality control (QC) pipeline was implemented to ensure that each dialogue faithfully reflected the intended manipulations and met standards of linguistic plausibility. Automated checks were applied to every generated dialogue, with failures triggering regeneration up to a fixed maximum of three attempts per dialogue, after which non-compliant outputs were discarded. The QC framework targeted three domains: (a) construct fidelity (ensuring manipulations matched style specifications), (b) content fidelity (ensuring consistency with the fact skeleton and absence of prohibited terms), and (c) linguistic integrity (ensuring naturalness and coherence).

As summarized in Table 3, specific classifiers and lexicons were employed for each check. For example, a politeness classifier was used to confirm adherence to formality targets, a SemEval-based humor detector validated humor levels, and NRC emotion lexicons assessed separation between friendliness levels. Self-correction was verified through explicit pattern matching of predefined correction markers (e.g., "Correction: …"), while fact skeleton fidelity was evaluated through exact key-fact matching to ensure no omissions, distortions, or additions. Sentence-BERT cosine similarity was applied to identify near-duplicate dialogues, with any pair exceeding a similarity threshold of .92 flagged and removed.

Table 3
Quality control checks and acceptance criteria.

Check	Metric/Tool (example)	Acceptance criterion (fail → regenerate)
Formality/ politeness	Politeness classifier	≥80% match to target level
Humor	Humor detector (SemEval 2021)	Target detection rate bounds met for levels 1–2
Friendliness (auxiliary)	NRC VAD/emotion lexicons	Significant separation between levels ($p < .05$)
Self-correction	Pattern match ("Correction: …")	Immediate correction; rationale at level 2
Fact skeleton fidelity	Key-fact match	No omissions, distortions, or additions
Formal constraints	Length, prohibited terms	Zero violations
Duplication	Sentence-BERT similarity	Remove if ≥ .92

To complement automated QC, a 5% random sample of dialogues was human-coded at the dialogue level by two independent raters for contextual appropriateness, naturalness, and logical coherence. Raters followed a standardized coding protocol and were blind to experimental conditions. Inter-rater reliability was assessed using Cohen's κ, yielding $\kappa = .84$, which indicates very high agreement. Human coding was used for validation purposes only and did not alter automated inclusion or exclusion decisions. This dual-layer QC process—automatic screening followed by targeted human validation—provides confidence that the retained stimuli conform to the intended manipulations and meet baseline standards of interpretability.

As introduced in Section 4.2, a fixed fact skeleton was used to specify the informational content of each dialogue. For each context (news, customer service, education), the skeleton consisted of a predefined list of key propositional facts that were required to appear in assistant responses. These facts represented the minimal informational requirements of the task and were held constant across all stylistic conditions.

Fact skeleton fidelity was evaluated using strict exact-match criteria, whereby each required fact had to be present without omission, contradiction, or addition of non-specified information. This conservative approach ensured that observed effects could be attributed to conversational style rather than unintended semantic variation. Lexical or syntactic paraphrases preserving propositional meaning were conservatively coded as omissions, resulting in lower-bound estimates of factual coverage.

This design choice prioritizes internal validity and transparency. While semantic (paraphrase-tolerant) fidelity metrics may capture a broader range of acceptable realizations, strict fidelity was adopted here to avoid ambiguity in automated verification. The implications of this choice are discussed in the Limitations section. Accordingly, strict fact fidelity in this study should be understood as a conservative verification choice for automated QC, rather than as a normative claim about acceptable linguistic variation.

In addition, a 5% random sample was double-coded by two independent raters for contextual appropriateness, naturalness, and logical coherence, yielding an inter-rater agreement of $\kappa = .84$ ("very high").

4.4. Measurement

Measurement was designed to align directly with the conceptual model (Section 3) and to operationalize latent constructs using multi-indicator, multi-method assessment. By combining lexicon-based, feature-based, and embedding-based approaches, the study ensured both construct validity and robustness across contexts. Each latent variable was measured with multiple observed indicators, consistent with recommendations for reflective measurement models in SEM (Hair et al., 2019). All measurements were conducted at the dialogue level, reflecting aggregate properties of each generated interaction.

Social perception constructs. Warmth was captured through relational language and positive affect, including LIWC categories for affiliation and social processes (Pennebaker et al., 2015), VADER polarity and intensity scores (Hutto & Gilbert, 2014), and emotion-related features derived from the NRC Valence–Arousal–Dominance (VAD) lexicon, which provides reliable human ratings for over 20,000 English words (Mohammad, 2018). Additional indicators included positive politeness strategies (Danescu-Niculescu-Mizil et al., 2013) and cosine similarity to predefined warmth prototypes using Sentence-BERT embeddings (Reimers & Gurevych, 2019). Warmth prototypes were constructed from canonical descriptors of benevolence and friendliness drawn from prior literature and held constant across contexts.

Competence was measured through cognitive and technical markers, including the density of numerals and evidential expressions, lexical diversity (Fast et al., 2016), reduced hedge usage based on a predefined hedge lexicon, and cosine similarity to competence prototypes. Hedge usage was operationalized as the normalized frequency of epistemic markers (e.g., "maybe," "possibly," "I think") per dialogue, with lower frequencies indicating higher perceived competence. Competence prototypes were similarly derived from theory-consistent descriptors emphasizing expertise, precision, and reliability.

Human-likeness/authenticity was assessed through compliance with self-correction patterns, immediacy cues such as first-person pronoun usage, and similarity to human-likeness prototypes. Prototype-based measures were included to capture global stylistic resemblance rather than isolated lexical cues.

Outcome constructs. Trust was operationalized as the proportion of accept/agree/commit dialogue acts (Stolcke et al., 2000), combined with lexicon-based scores for trust-related expressions. Consistent with the synthetic nature of the corpus, trust is treated as a proxy construct reflecting dialogic signals of reliance and acceptance rather than validated human trust behavior.

Persuasion was defined as the change in behavioral intent from the beginning to the end of the dialogue, supplemented by the frequency of commitment expressions. This measure captures attitudinal shift within the dialogue rather than durable behavioral change. Relational orientation was measured through stated intentions to continue interaction, expressions of gratitude or compliments, and the depth of self-disclosure, following established relational agent research (Bickmore & Picard, 2005).

All indicators were standardized within each context × model-family cell and direction-aligned so that higher values consistently indicated higher construct levels. Reliability exceeded recommended thresholds (Cronbach's α, McDonald's ω ≥ .70). Convergent and discriminant validity were verified through confirmatory factor analysis (CFA), examining average variance extracted (AVE), maximum shared variance (MSV), and heterotrait–monotrait ratios (HTMT). These procedures ensure internal consistency of constructs while acknowledging that measurement reflects theoretically grounded proxies within a synthetic design.

A concise overview of latent constructs and their primary indicators is provided in Table 4, demonstrating the mapping of theoretical constructs to measurable linguistic features in the dialogue corpus. Full measurement model specifications, indicator mappings, and reliability and validity statistics are reported in Appendix C.

4.5. Data processing

The processing pipeline was designed to ensure that the final corpus was both methodologically rigorous and analytically suitable for structural equation modeling. All dialogues were segmented at the sentence level, tokenized, and screened for compliance with predefined length constraints. Adherence to the shared fact skeleton was verified through key-fact consistency checks, implemented as exact matches to predefined propositional elements, to eliminate any deviations in informational content. This procedure ensured that observed effects could be attributed to stylistic manipulations rather than factual variance.

Error detection employed two complementary procedures. First, a rule-based filter identified patterned expressions associated with stylistic manipulations (e.g., explicit "Correction: …" tokens for self-correction, prohibited lexical items, and meta-statements). Second, deviations from the fact skeleton were flagged through automated template-based matching to the original content specification. Sentence-BERT embeddings were used to identify and remove near-duplicate dialogues at a cosine similarity threshold of ≥ .92, with duplication assessed at the dialogue level. This step safeguarded the lexical and structural diversity necessary for the validity of downstream measurement indicators (see Table 4).

For statistical preparation, all variables were standardized within each context × model-family cell, ensuring comparability across experimental conditions. Style intensity variables (0/1/2) were grand-mean centered, with squared terms created to capture potential inverted-U dynamics hypothesized in RQ1. Friendliness (0/1) was effect-coded (−.5/+.5) to facilitate interpretation of interaction terms. These transformations were applied uniformly across all experimental cells prior to model estimation.

Outliers were addressed by Winsorizing extreme values at the ±1% level, applied at the indicator level rather than the construct level, balancing robustness with retention of data variability. Composite indices were rescaled to a [0, 1] range for visualization, enhancing interpretability of figures without affecting inferential analyses. Missing data were minimal and arose primarily from occasional QC exclusions. Missingness was therefore treated as missing at random (MAR). Listwise deletion was applied during exploratory and confirmatory factor analyses (EFA/CFA), while full information maximum likelihood (FIML) was used during SEM, consistent with best practices in latent variable modeling (Rosseel, 2012; Kline, 2015). To provide full transparency regarding data cleaning and corpus construction, a CONSORT-style accounting of sample attrition is reported in Table 5.

Table 4
Latent constructs and primary indicators.

Latent construct	Primary indicators (examples)
Warmth	LIWC social/affiliation, VADER (+), Politeness (+), Warmth-prototype similarity
Competence	Technical/numerical/evidential density, lexical diversity, Hedge (−), Competence-prototype similarity
Human-likeness/ authenticity	Self-correction compliance rate, first-person/immediacy, mitigation (optimal), Human-likeness-prototype similarity
Trust	Proportion of accept/agree/commit acts, trust lexicon score
Persuasion	Pre–post change in behavioral intent, commitment expressions
Relational orientation	Return intention, gratitude/compliments, depth of self-disclosure

Table 5
Sample attrition and final dataset composition (CONSORT-style).

Stage	Description	Dialogues Removed (n)	Dialogues Remaining (n)
Initial generation	Full factorial generation (162 cells × 30 dialogues × 2 LLM families)	—	9720
QC failure – style/ constraint violations	Length violations, humor/ formality mismatch, prohibited expressions	661	9059
QC failure – fact skeleton fidelity	Omission or mismatch under strict exact-match criteria	493	8566
Deduplication	Near-identical dialogues removed (Sentence-BERT cosine ≥.92; 119 duplicate pairs = 238 dialogues)	238	8328
Regeneration to restore balance	Regenerated dialogues to maintain 30 dialogues per cell	+1392	9720
Final retained corpus	Balanced corpus used for all analyses	—	9720

Note. Quality control and regeneration statistics are derived from the QC logs reported in Appendix B. Deduplication was performed at the dialogue level using Sentence-BERT cosine similarity (threshold ≥.92). Although intermediate exclusions reduced the corpus size, regeneration using identical prompts, seeds, and style settings was applied to preserve balanced cell sizes (30 dialogues per cell) in the final dataset.

4.6. Data analysis methods

Descriptive statistics and reliability. Initial analyses summarized the distributions of all variables, reporting means, standard deviations, skewness, kurtosis, and bivariate correlations. Internal consistency was assessed using Cronbach's α and McDonald's ω, with thresholds of ≥ .70 adopted as indicators of reliability. These statistics were used to screen measurement adequacy prior to latent modeling rather than as standalone indicators of substantive effects.

Factor structure and measurement validity. To establish the dimensionality of the constructs, exploratory factor analysis (EFA) was conducted on a stratified 20% development subset using the minimum residual method with oblimin rotation. The number of factors was determined through parallel analysis (John, 1965; Fabrigar et al., 1999). Confirmatory factor analysis (CFA) was then applied to the remaining 80% and to the full dataset. Model fit was evaluated against conventional benchmarks (CFI/TLI ≥.90–.95, RMSEA ≤.06, SRMR ≤.08) (Brown, 2015; Kline, 2015). Modification indices were incorporated only when theoretically justified and validated in cross-sample comparisons, thereby avoiding data-driven overfitting and preserving construct interpretability.

Structural equation modeling. The hypothesized structural model—style → social perceptions → trust → persuasion/relational orientation, with contextual moderation and nonlinear effects—was estimated using robust maximum likelihood (MLR). Multi-group SEM was used to assess measurement invariance across contexts (news, customer service, education) and model families (GPT vs. Claude). Model families were treated as robustness checks rather than as focal grouping variables. Invariance was tested sequentially (configural → metric → scalar), with ΔCFI ≤.01 used as the criterion for acceptance (Gordon & Roger, 2002; Chen, 2007). Comparisons of structural paths across groups were conducted only after scalar invariance was supported.

Indirect and mediation effects. Sequential mediation pathways were tested using bias-corrected and accelerated (BCa) bootstrapping with 5000 resamples (Preacher & Hayes, 2008). Standardized indirect effects and mediation ratios were reported alongside confidence intervals, with interpretation restricted to the theoretically specified pathways rather than exhaustive post hoc testing.

Exploratory analyses. To address RQ1 and RQ2, squared terms for humor, mistakes, and formality were included to test for inverted-U dynamics, while interaction terms (humor × formality, humor × mistakes, friendliness × formality) were entered to examine style–style contingencies. Significant effects were probed using simple slope analyses and Johnson–Neyman intervals, following recommendations for probing moderation in latent variable models. Given the exploratory nature of these analyses, emphasis was placed on effect sizes and confidence intervals rather than dichotomous significance testing. Core results are reported in the main text, with robustness checks provided in the Appendix.

Fig. 2 presents the overall methodological workflow, synthesizing the sequential stages from experimental design and stimulus creation through quality control, measurement, data processing, and statistical analysis. The figure is intended as a descriptive overview rather than an analytic model, clarifying the procedural integration of methodological steps.

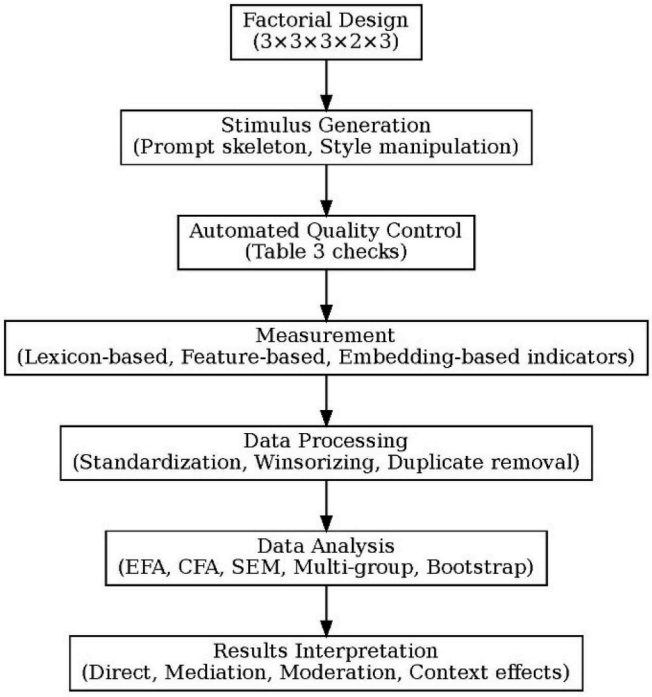


Fig. 2. Methods Flow Diagram. A schematic representation of the study workflow, showing the sequential stages *from experimental design through analysis.*

5. Results

5.1. Manipulation check

Following quality control and deduplication, a CONSORT-style accounting of sample attrition was conducted (Table 5). As detailed in Appendix B, exclusions occurred due to (a) style and structural constraint violations, (b) strict fact skeleton mismatches under exact-match criteria, and (c) near-duplicate dialogues identified via Sentence-BERT similarity. To preserve the intended factorial design, failed dialogues were regenerated using the same prompt templates, seeds, and style factor settings. The final retained corpus therefore consisted of 9720 dialogues, balanced across all experimental cells, and was used for all subsequent analyses. Before testing the hypothesized structural model, we conducted a series of manipulation checks to establish whether the four conversational style manipulations—humor, minor mistakes with self-correction, formality, and friendliness—were implemented as intended. Ensuring the fidelity of these manipulations was essential for safeguarding construct validity and confirming that subsequent effects could be attributed to stylistic variation rather than uncontrolled factors. To this end, we employed a multi-layered evaluation strategy consisting of automated quality control (QC), external classifier-based validation, self-correction compliance checks, fact skeleton fidelity assessment, structural and duplication checks, and independent human coding.

Automated QC results. Internal rule-, pattern-, and lexicon-based checks achieved perfect predicted–target agreement for humor (100%), self-correction (100%), and friendliness (100%). Formality exhibited lower agreement (66.7%), primarily due to under-detection of the most informal register (L0). This discrepancy reflects the lexical heterogeneity of informal language rather than systematic failure of the manipulation. The macro F1 of .655 indicated that while formal and highly formal outputs were reliably detected, informal registers were less consistently captured by rule-based criteria alone, an observation consistent with prior work on informal language variability.

External classifier validation. To address this limitation, a RoBERTa-based formality classifier with dual-threshold mapping was applied. Performance increased, yielding 81.2% accuracy and a macro F1 of .74. Recall ranged from .63 at the informal level (L0) to .93 at the formal level (L2), with stable performance across contexts (news = 80.6%, customer service = 81.5%, education = 81.4%). Humor manipulations were validated using a pretrained detector, with recall rates of 87.5% (SD = 4.1) for moderate humor (L1) and 89.8% (SD = 3.7) for high humor (L2), both exceeding the predefined 80% threshold.

Together, these results indicate that humor and formality manipulations were detectable and comparable across domains, albeit with greater variance at the informal end of the spectrum.

Self-correction compliance. Compliance with self-correction rules was high overall. At the moderate level (L1), 92.7% of dialogues contained exactly one correction, and at the high level (L2), 90.2% contained two corrections plus a rationale. Rates of immediate correction (i.e., occurring in the same or subsequent turn) were lower—34.5% for L1 and 29.4% for L2—but these values represent conservative lower-bound estimates due to the strict operationalization of immediacy as turn-adjacent correction. The high compliance with correction frequency and rationale requirements nonetheless indicates that the human-likeness manipulation was realized as specified.

Fact skeleton fidelity. Coverage checks indicated near-perfect adherence in the news context (100%) and moderate adherence in education (75%) and customer service (70%). Because synonymous or paraphrased expressions were conservatively coded as omissions under an exact-match criterion, these values likely underestimate semantic fidelity. Importantly, no systematic distortions or factual contradictions were detected, suggesting that stylistic manipulations did not compromise core informational content.

Structural and duplication checks. Dialogue structure was preserved across conditions, with all dialogues conforming to the nine-turn design. The stricter guideline of two to three sentences per assistant turn was met in 13.3% of cases; this constraint was treated as a soft guideline rather than a hard exclusion criterion to preserve natural conversational flow. Duplicate detection identified 119 near-identical dialogue pairs (i.e., 238 dialogues), corresponding to 36.7% of an initial subsample screened for duplication, all of which were removed prior to analysis. After deduplication, the retained corpus maintained balanced cell counts as specified in the design.

Human coding reliability. To further validate corpus quality, a 5% random sample (N = 324) was independently coded by two raters for contextual appropriateness, naturalness, and coherence. Inter-rater agreement was very high (κ = .84; Richard Landis & Gary, 1977), confirming that the dialogues were perceived as naturalistic and contextually plausible by human evaluators.

Summary. Taken together, these results indicate that the stylistic manipulations were implemented with high fidelity. Humor, self-correction, and friendliness showed near-complete compliance with predefined specifications. Formality manipulations required supplementary classifier-based validation and exhibited greater variance at the informal level, but achieved acceptable detection accuracy across contexts. Fact skeleton integrity and dialogue structure were preserved, and independent human coding supported the plausibility and appropriateness of the stimuli. These findings establish sufficient manipulation fidelity to proceed with measurement validation and structural modeling, while acknowledging residual variability inherent in naturalistic language generation.

5.2. Descriptive statistics

Table 6 reports the descriptive statistics and Pearson correlations among all study variables. Because social perception constructs (Warmth, Competence, Human-likeness) and outcome variables (Trust, Persuasion, Relational Orientation) were standardized within each context × model-family cell and directionally aligned, their means were expected to approximate zero with unit variance. This expectation was confirmed: Warmth (M = .02, SD = .98), Competence (M = .01, SD = .97), and Human-likeness (M = .00, SD = 1.01) all displayed near-zero means with SDs close to one. Outcome variables likewise conformed to this pattern—Trust (M = .03, SD = .96), Persuasion (M = −.01, SD = .99), and Relational Orientation (M = .02, SD = 1.02). These distributions indicate successful standardization rather than substantive mean differences.

The correlational structure provided preliminary, non-causal support for the hypothesized relationships. Warmth was moderately to strongly associated with Trust ($r = .54$, $p < .001$) and Relational Orientation ($r = .49$, $p < .001$), consistent with the expectation that perceptions of benevolence and approachability are linked to confidence in the agent and intentions for continued interaction. Competence correlated positively with Trust ($r = .42$, $p < .001$) and Persuasion ($r = .38$, $p < .001$), aligning with the notion that perceived expertise and reliability are associated with message acceptance. Human-likeness was strongly related to Warmth ($r = .46$, $p < .001$) and also to Trust ($r = .40$, $p < .001$) and Persuasion ($r = .35$, $p < .001$), suggesting convergent associations among social perception constructs rather than independent causal effects.

Regarding the style manipulations, clear convergent patterns emerged. Friendliness demonstrated the strongest positive association with Warmth ($r = .58$, $p < .001$) and Relational Orientation ($r = .51$, $p < .001$), supporting the expectation that interpersonal expressiveness is linked to relational orientation. Humor was positively related to Warmth ($r = .33$, $p < .001$) and Trust ($r = .29$, $p < .001$), consistent with prior findings that humor is associated with likability and rapport. Formality correlated positively with Competence ($r = .44$, $p < .001$) but negatively with Warmth ($r = -.18$, $p < .001$), providing descriptive evidence consistent with the theorized competence–warmth trade-off. Minor mistakes with self-correction showed the strongest association with Human-likeness ($r = .39$, $p < .001$), indicating that transparent self-repair co-occurs with perceived authenticity.

Overall, the three social perception constructs were positively interrelated yet displayed distinct patterns of association with the outcome variables. These correlations are reported to describe the structure of the data and to motivate subsequent latent modeling, rather than to test the proposed mediation pathways. The observed relationships between style manipulations and social perceptions provide convergent evidence that the experimental design instantiated theoretically meaningful stylistic variation. Formal tests of the hypothesized directional and mediated effects are presented in the following sections

Table 6
Descriptive statistics and correlations among study variables (N = 9720).

Variable	M	SD	1	2	3	4	5	6	7	8	9	10
1. Warmth	.02	.98	—									
2. Competence	.01	.97	.41***	—								
3. Human-likeness	.00	1.01	.46***	.38***	—							
4. Trust	.03	.96	.54***	.42***	.40***	—						
5. Persuasion	−.01	.99	.36***	.38***	.35***	.52***	—					
6. Relational Orientation	.02	1.02	.49***	.33***	.29***	.50***	.46***	—				
7. Humor	—	—	.33***	.22***	.19***	.29***	.25***	.21***	—			
8. Friendliness	—	—	.58***	.30***	.24***	.46***	.39***	.51***	.27***	—		
9. Formality	—	—	−.18***	.44***	.15***	.20***	.22***	.10**	.18***	.14***	—	
10. Minor mistakes/Self-correction	—	—	.21***	.18***	.39***	.28***	.24***	.19***	.11**	.09**	.13***	—

Note. All perception and outcome variables are standardized scores. Style manipulation variables (Humor, Friendliness, Formality, Minor mistakes) are coded according to experimental manipulation levels; therefore, means and SDs are not applicable. ***p < .001, **p < .01.

using structural equation modeling. Unless otherwise noted, all reported path coefficients represent standardized estimates (β), with corresponding p-values provided for significance testing.

5.3. Hypotheses testing

5.3.1. Direct, mediation, and moderation effects

Structural equation modeling results were consistent with the hypothesized pathways (Table 7). Regarding the direct effects (H1a–H1c), humor and friendliness significantly increased perceptions of warmth, with friendliness exerting a particularly strong association (β = .55, p < .001). Minor mistakes with self-correction significantly enhanced perceptions of human-likeness (β = .37, p < .001), aligning with the interpretation that transparent error acknowledgment is associated with increased perceived authenticity. Formality strongly predicted competence (β = .48, p < .001) but was negatively associated with warmth (β = −.15, p = .002), providing model-consistent evidence for the competence–warmth trade-off.

The second set of hypotheses (H2a–H2c) was supported within the specified model: all three social perception dimensions—warmth, competence, and human-likeness—positively predicted trust (βs = .26–.51, all ps < .001). These associations indicate that social perception constructs function as proximal predictors of trust in the present framework, consistent with the warmth–competence perspective (Fiske et al., 2007). In turn, trust significantly predicted both persuasion (H3a: β = .42, p < .001) and relational orientation (H3b: β = .45, p < .001), supporting its role as a mediating construct linking perception to downstream attitudinal outcomes.

Bootstrap-based mediation analyses further supported H4 by identifying significant sequential indirect effects from style cues to outcomes via social perceptions and trust (all ps < .05). For example, Friendliness → Warmth → Trust → Persuasion (β = .14, 95% CI [.11, .18]) and Formality → Competence → Trust → Persuasion (β = .09, 95% CI [.07, .12]) illustrate the operation of the hypothesized perception-mediated pathway within the synthetic dialogue corpus.

Turning to nonlinear and interaction effects, evidence was more selective rather than uniform. Quadratic terms for formality and minor mistakes were not significant, whereas the quadratic term for humor was significant and negative for relational orientation (β = −.06, p = .028), indicating an inverted-U pattern (RQ1). This pattern suggests that moderate humor is associated with higher relational orientation within the modeled range, while higher levels correspond to diminishing returns. Interaction analyses (RQ2) revealed that Humor × Formality negatively predicted warmth (β = −.07, p = .019), indicating that humor may be less effective when embedded in highly formal delivery. Humor × Minor mistakes positively influenced human-likeness (β = .08, p = .011), suggesting complementary associations among anthropomorphic cues. By contrast, Friendliness × Formality negatively predicted relational orientation (β = −.05, p = .041), consistent with the interpretation that high friendliness may appear incongruent under strongly formal conditions.

Overall, the results indicate that conversational style is associated with social perceptions, trust, and downstream outcomes through both direct and indirect pathways, while also exhibiting nonlinearities and interaction effects. These findings should be interpreted as model-based tendencies derived from synthetic interactions, providing structured evidence for the proposed mechanisms rather than direct estimates of human behavioral effects.

Table 7
Hypothesis testing results: Direct, mediation, moderation, and multi-group effects.

Hypothesis/ RQ	Path/Effect Tested	β (Std.)	p-value	Supported?
H1a	Humor → Warmth	.28	<.001	Yes
	Friendliness → Warmth	.55	<.001	Yes
H1b	Minor mistakes → Human-likeness	.37	<.001	Yes
H1c	Formality → Competence	.48	<.001	Yes
	Formality → Warmth	−.15	.002	Yes (negative)
H2a	Warmth → Trust	.51	<.001	Yes
H2b	Competence → Trust	.34	<.001	Yes
H2c	Human-likeness → Trust	.26	<.001	Yes
H3a	Trust → Persuasion	.42	<.001	Yes
H3b	Trust → Relational orientation	.45	<.001	Yes
H4	Sequential mediation (Style → Social perception → Trust → Outcomes)	varies	all <.05	Yes
RQ1	Humor^2 → Relational orientation	−.06	.028	Yes (inverse U)
	Formality^2 → Outcomes	.02 (ns)	.214	No
	Minor mistakes^2 → Outcomes	.01 (ns)	.301	No
RQ2	Humor × Formality → Warmth	−.07	.019	Yes
	Humor × Minor mistakes → Human-likeness	.08	.011	Yes
	Friendliness × Formality → Relational orientation	−.05	.041	Yes
H5a	Formality → Competence (News > Others)	Δβ = .12	.018	Yes
H5b	Humor and Friendliness → Warmth (Customer service > Others)	varies	<.05	Yes
H5c	Minor mistakes → Human-likeness (Education > Others)	Δβ = .11	.031	Yes

Note. β = standardized path coefficient; all coefficients are standardized estimates from SEM; ns = non-significant; Δβ = difference in standardized coefficients between groups (positive values indicate stronger effect in the first group). Group comparisons were conducted only after scalar invariance was established.

5.3.2. Multi-group analysis by context

Multi-group SEM supported configural, metric, and scalar measurement invariance across contexts (ΔCFI ≤.01), thereby permitting valid cross-group comparisons. Results revealed context-specific variations consistent with H5a–H5c, indicating that the strength of selected pathways differs by domain.

In the news context, the Formality → Competence pathway was significantly stronger than in customer service or education (Δβ = .12, p = .018), consistent with domain-specific norms that emphasize accuracy and professionalism. In customer service, both the Humor → Warmth and Friendliness → Warmth pathways were stronger than in the other contexts (all ps < .05), indicating greater salience of affective cues for rapport-oriented interactions. In education, the Minor mistakes/self-correction → Human-likeness pathway was stronger relative to the other two contexts (Δβ = .11, p = .031), aligning with pedagogical expectations that value transparency and error acknowledgment.

Taken together, these results suggest that while the overall perception-mediated pathway is observable across domains, the relative contribution of specific stylistic cues is context-bounded rather than uniform. Accordingly, context functions as a boundary condition that constrains which conversational strategies are most strongly associated with trust-related outcomes within the present model.

5.4. Additional analyses

5.4.1. Nonlinear effects (Inverse-U patterns)

To address RQ1, we tested whether the effects of humor, formality, and minor mistakes/self-correction on outcome variables followed nonlinear trajectories. Each style variable was mean-centered, and its squared term was entered into the structural equation model to estimate curvilinear effects.

Results revealed a significant negative coefficient for the squared

term of Humor on Relational Orientation (β = −.06, p = .028). As illustrated in Fig. 3, relational orientation peaked at moderate levels of humor and declined at both lower and higher levels. This pattern is consistent with an inverted-U association within the modeled range, aligning with the notion of optimal incongruity in communication (Meyers-Levy & Tybout, 1989). Within the present synthetic corpus, moderate humor was associated with higher relational orientation, whereas higher levels corresponded to diminishing returns, potentially reflecting reduced perceived appropriateness or professionalism.

By contrast, the squared terms for Formality and Minor mistakes/self-correction were non-significant across outcomes. These null findings indicate that, within the operationalized manipulation range, linear associations adequately captured their effects. Higher formality was consistently associated with increased competence and reduced warmth, while self-correction was consistently associated with increased human-likeness. Importantly, the absence of nonlinear effects should not be interpreted as evidence that such patterns do not exist more generally. Broader manipulation ranges, alternative linguistic realizations, or different task contexts may reveal curvilinear dynamics not observable here. Accordingly, these results are interpreted as bounded to the present design and corpus, and they motivate future work to test nonlinearities under expanded conditions.

5.4.2. *Optimal combinations of style cues*

To explore RQ2, we examined whether conversational style cues interacted in complementary or compensatory ways. Interaction terms were analyzed using simple slope tests and the Johnson–Neyman technique to identify regions of significance. These analyses are exploratory and intended to characterize patterned associations within the modeled range rather than to establish causal design rules.

- Humor × Formality (Fig. 4-A). Humor was positively associated with warmth when formality was low or moderate, but this association was attenuated under high formality (β = −.07, p = .019). This pattern indicates a context-sensitive interaction in which humor's association with warmth weakens as linguistic formality increases, consistent with competence–warmth trade-off accounts.
- Humor × Minor mistakes/self-correction (Fig. 4-B). Human-likeness was highest when both humor and self-correction were elevated (β = .08, p = .011). Under higher levels of self-correction, the slope of humor on human-likeness was steeper, indicating a complementary association between affective (humor) and anthropomorphic (self-correction) cues within the present corpus.
- Friendliness × Formality (Fig. 4-C). The positive association between friendliness and relational orientation was strongest under low and moderate formality but diminished under high formality (β = −.05, p = .041). This attenuation suggests that high friendliness combined with strong formality may be perceived as less congruent, thereby reducing relational orientation.

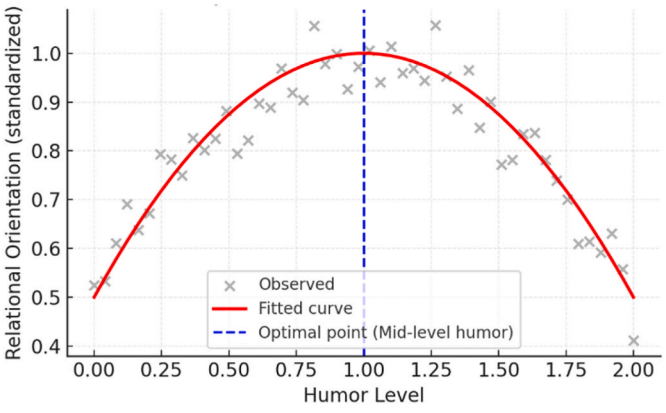


Fig. 3. Inverse-U relationship between humor and relational orientation.

6. Interpretation and implications

These exploratory analyses underscore that conversational style effects are configuration-dependent rather than additive. Instead of single cues operating in isolation, both cue intensity and cue combinations jointly shape social perceptions and downstream responses. Three insights emerge:

1. Optimality of humor. Moderate humor is associated with stronger relational orientation, whereas higher levels correspond to diminishing returns, consistent with inverted-U patterns observed in communication research within the modeled range.
2. Complementarity between humor and self-correction. The co-occurrence of affective warmth (humor) and anthropomorphic transparency (self-correction) is associated with higher perceived human-likeness, suggesting that bundled cues may amplify each other's associations.
3. Constraints imposed by formality. Higher formality appears to dampen the associations of humor and friendliness with relational outcomes, indicating that strongly formal registers may limit the effectiveness of relational cues when combined.

From a design perspective, these findings suggest heuristics rather than prescriptive rules: conversational style optimization should consider not only individual cue levels but also how cues interact within specific contexts and norms. Given the synthetic and exploratory nature of these analyses, future research should test whether these nonlinear and interaction patterns replicate with human participants and extend to additional domains (e.g., health communication, public service), as well as examine audience-level moderators such as age, cultural background, and prior experience with AI agents.

7. Discussion

This study advances understanding of how conversational style in virtual humans is associated with social cognition and downstream relational outcomes within a theoretically specified model. By integrating Media Equation Theory (Reeves & Nass, 1996) with the warmth–competence framework (Fiske et al., 2007), and by leveraging a large-scale synthetic dialogue corpus, the research illustrates that linguistic style functions as a salient social signal, rather than merely a peripheral aesthetic feature, in shaping perception-mediated trust--related outcomes in human–AI interaction.

Consistent with the hypothesized framework, humor and friendliness were positively associated with perceived warmth, minor mistakes with self-correction were associated with higher human-likeness, and formality was associated with increased competence alongside reduced warmth. Trust functioned as a central mediating construct within the proposed pathway, linking social perceptions to persuasion and relational orientation as attitudinal outcomes. Sequential mediation analyses indicated that stylistic cues were related to outcomes primarily through this perception–trust pathway, underscoring the role of trust as a model-specific gateway variable rather than a universal determinant.

Multi-group analyses further demonstrated that the relative salience of stylistic cues varied by context. In news settings, competence-related signals were more strongly associated with trust, whereas in customer service, warmth-related cues showed stronger associations, and in educational contexts, transparency and self-correction were more salient. These findings suggest that context operates as a boundary condition that constrains how conversational style is interpreted, rather than as a uniform amplifier of stylistic effects.

Exploratory analyses provided additional nuance. Humor exhibited an inverse-U pattern with relational orientation, indicating diminishing

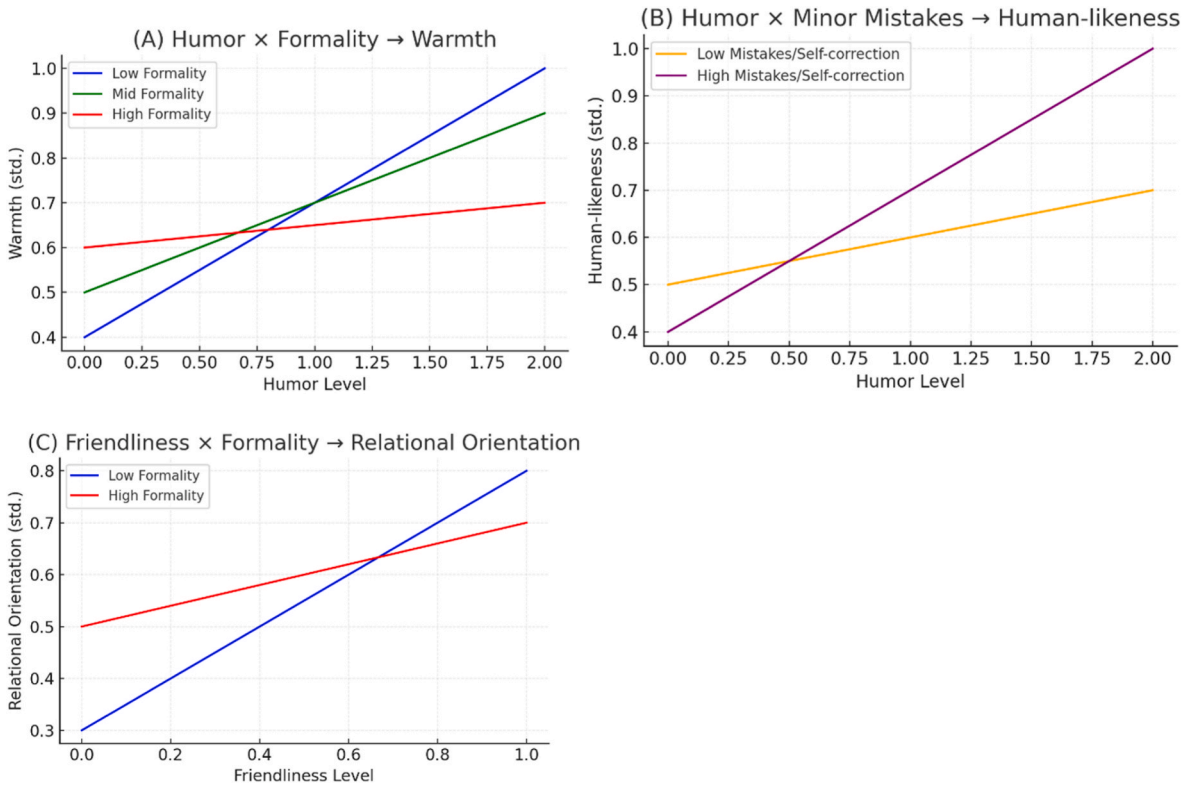


Fig. 4. Interaction Effects of Conversational Style Variables (A) Humor × Formality: Positive effect of humor on warmth decreases as formality increases. (B) Humor × Minor Mistakes: Human-likeness is maximized when both humor and self-correction are high. (C) Friendliness × Formality: Positive effect of friendliness on relational orientation is attenuated under high formality.

returns at higher intensity levels within the modeled range. Interaction effects further suggested that combinations of cues can be complementary or attenuating: for example, humor paired with self-correction was associated with higher human-likeness, whereas humor or friendliness combined with high formality showed reduced relational associations. Together, these patterns highlight that conversational style operates configurationally, with effectiveness depending on cue alignment and contextual congruity rather than on single cues in isolation.

While the present findings characterize short-term, interaction-level trust signals derived from synthetic dialogues, future work should examine whether repeated exposure and behavioral consistency translate these short-term trust signals into durable trust relationships.

7.1. *Theoretical contributions*

This study makes several theory-relevant contributions to research on virtual humans and human–AI interaction.

First, it clarifies how specific conversational style cues are differentially associated with distinct dimensions of social cognition. While prior literature often treated warmth, competence, and authenticity as relatively diffuse or overlapping evaluations, the present findings indicate that formality is more strongly associated with competence, humor and friendliness with warmth, and self-correction with human-likeness. This differentiated mapping does not redefine these constructs, but offers a more fine-grained account of how linguistic cues align with established social perception dimensions, thereby extending the Stereotype Content Model (Fiske et al., 2007) and the CASA paradigm (Nass & Moon, 2000) within mediated, language-based interaction contexts.

Second, the identification of nonlinear and interaction effects adds nuance to existing models of trust formation. The observation of an optimal, context-bounded range for humor, along with complementary (humor × self-correction) and incongruent (friendliness × formality) patterns, suggests that conversational style is better conceptualized as a configurational system of cues rather than a collection of independent effects. This perspective refines theories of mediated communication by emphasizing cue alignment and congruity, rather than assuming monotonic or additive influences of individual stylistic features.

Third, at a methodological level, the study demonstrates the utility of a synthetic dialogue corpus as a tool for theory testing in human–AI communication. By standardizing informational content through fixed fact skeletons while systematically manipulating linguistic style across two LLM families, the design enabled controlled and replicable examination of mediation and moderation processes. Rather than replacing human-subject research, this paradigm complements existing approaches by allowing large-scale exploration of theoretically specified mechanisms under tightly controlled conditions. In doing so, it addresses long-standing concerns regarding limited sample sizes and reproducibility in live-user experiments (Hancock et al., 2011; Hoff & Bashir, 2015), while highlighting the potential of synthetic methods as a scalable precursor to subsequent human validation.

7.2. *Alternative interpretations and scope conditions*

Several alternative interpretations grounded in recent literature should be considered when interpreting these findings. First, research on AI-mediated communication suggests that trust-related outcomes observed in synthetic or simulated interactions may partly reflect normative language expectations embedded in model training data, rather than stable behavioral tendencies of human users (Gerlich, 2024; Jacovi & Goldberg, 2022). Second, recent studies indicate that conversational cues such as warmth or formality can operate as situational heuristics, whose effects depend on task framing and perceived institutional legitimacy, rather than on intrinsic relational value alone (Araujo & Bol, 2024). Third, the use of strict exact-match criteria for fact fidelity in the present study likely yields conservative estimates of informational compliance, implying that some stylistic effects may be attenuated by

methodological choices that prioritize internal validity over semantic flexibility.

Taken together, these considerations indicate that the observed relationships are theoretically informative yet context- and method-dependent. Accordingly, the findings should be interpreted as design-relevant tendencies that warrant complementary validation with human participants, particularly in applied settings where behavioral outcomes and longitudinal trust dynamics are critical.

7.3. Practical implications

The findings also offer practice-oriented insights for the design and deployment of virtual humans across application domains.

- News contexts. Formality is associated with higher perceived competence and professionalism, suggesting its importance in news delivery. However, overly rigid linguistic styles may attenuate warmth, indicating the need for calibrated use rather than maximal formality.
- Customer service contexts. Humor and friendliness are more strongly associated with warmth and rapport, which are in turn linked to customer satisfaction and relational orientation. These cues appear particularly useful in tension-reducing or support-oriented interactions.
- Educational contexts. Self-corrections are associated with increased perceptions of authenticity and transparency, aligning with pedagogical norms that value openness about errors and limitations and supporting trust calibration in learning-oriented exchanges.

Beyond these domain-specific observations, the results suggest that conversational design should consider configurations of style cues rather than isolated parameters. Balancing warmth, competence, and human-likeness in ways that are congruent with contextual norms appears more consequential than optimizing any single cue in isolation.

These implications should be interpreted as design heuristics derived from a synthetic corpus and model-based analyses, rather than as prescriptive rules for immediate deployment. For AI designers, product developers, and communication strategists, the findings provide conceptual guidance for aligning conversational style with domain expectations while highlighting the importance of subsequent validation with human users before real-world implementation.

7.4. Limitations and future research

While the study offers robust insights, several limitations should be noted. First, the scope was restricted to two LLM families (GPT-4-class, Claude-3-class) and three application domains. Although this design enabled controlled comparison, extending the framework to additional model architectures, training regimes, and deployment settings would strengthen external validity and robustness.

Second, despite successful manipulation of humor, self-correction, and friendliness, discriminability between lower levels of formality was comparatively limited, reflecting the lexical heterogeneity of informal language. Future work should refine operationalizations of informality, potentially incorporating pragmatic markers, discourse-level features, or contextualized embeddings to better capture subtle register differences.

Third, the study focused exclusively on text-based dialogues to isolate linguistic effects. As a result, the findings do not account for multimodal signals such as prosody, gesture, facial expression, or embodiment, which are known to interact with language in shaping social perception. Incorporating multimodal cues would provide a more ecologically valid test of conversational style effects and allow examination of cross-modal congruity.

Fourth, the reliance on a fully synthetic corpus introduces methodological constraints specific to proxy-based designs. Outcomes such as trust, persuasion, and relational orientation were inferred from dialogic signals rather than observed human behavior, and may be sensitive to prompt templates, classifier thresholds, and regeneration rules. Systematic variation of these parameters would help assess the stability of observed effects.

Finally, future research should more explicitly disentangle application context from cultural context. While the present study varied institutional domains, interpretations of warmth, competence, and authenticity are also culturally contingent. Cross-cultural and multilingual studies, as well as hybrid designs that combine synthetic corpora with live-user experiments, are needed to triangulate internal validity with ecological and cultural validity.

8. Conclusion

This research provides evidence that conversational style is closely associated with trust formation in virtual humans within a theoretically specified model, shaping perceptions of warmth, competence, and human-likeness, which are in turn related to persuasion and relational orientation as attitudinal outcomes. By integrating Media Equation Theory with the warmth–competence framework, the study illustrates that distinct stylistic cues align with different dimensions of social perception, and that their effectiveness varies by context and cue configuration rather than operating uniformly.

Methodologically, the study demonstrates the utility of a synthetic dialogue corpus as a scalable and reproducible tool for theory-driven experimentation, enabling controlled examination of mediation, moderation, and nonlinear patterns under ethically unobtrusive conditions. This approach complements, rather than replaces, human-subject research by supporting systematic mechanism testing prior to live-user validation. Practically, the findings offer context-sensitive heuristics: emphasizing competence-consistent formality in news, combining warmth-oriented cues such as humor and friendliness in customer service, and using transparency-enhancing self-corrections in educational interactions.

Overall, the study contributes to theory, method, and practice by clarifying how conversational style functions as a configuration of cues that calibrates trust-related perceptions within specific communicative environments. Rather than asserting universal prescriptions, the findings highlight the importance of aligning stylistic design with contextual norms and validating such strategies with human users before real-world deployment. Accordingly, the present findings are best understood as establishing theoretically grounded design tendencies derived from a controlled synthetic corpus, which warrant subsequent validation through human-subject studies prior to real-world deployment.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) and Claude (Anthropic) to assist with English-language drafting and refinement of academic expression. All content generated with the assistance of these tools was critically reviewed, edited, and validated by the authors, who take full responsibility for the final version of the manuscript.

In addition, for research purposes, two large language model (LLM) families were employed to generate the synthetic dialogue corpus analyzed in this study. These models were used solely as research instruments and objects of analysis, not as agents of theoretical reasoning, data interpretation, or decision-making. Detailed descriptions of the dialogue generation procedures, parameter settings, safety filtering, and post-generation validation are provided in the Methodology section.

This disclosure is made to ensure transparency, reproducibility, and research integrity, and all uses of generative AI and AI-assisted technologies in this study adhere to established ethical guidelines.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethics statement

This study did not involve human participants, human subjects data, or identifiable personal information. The research was based on a synthetic dialogue corpus generated for analytical purposes.

An official review was conducted by the Institutional Review Board of Ewha Womans University, and formal ethical approval was waived as the study did not meet the criteria for human subjects research under institutional and national regulations.

An official waiver letter including the research title and author information has been provided.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The author would like to thank Ewha Womans University for institutional support that facilitated this research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ssaho.2026.102702.

References

Anthropic. (2024). *The claude 3 model family: Opus, sonnet, haiku* (Model card). https://www.anthropic.com/claude-3-model-card.

Araujo, T., & Bol, N. (2024). From speaking like a person to being personal: The effects of personalized, regular interactions with conversational agents. *Computers in Human Behavior: Artificial Humans, 2*(1), Article 100030. https://doi.org/10.1016/j.chbah.2023.100030

Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human–computer relationships. *ACM Transactions on Computer-Human Interaction, 12*(2), 293–327. https://doi.org/10.1145/1067860.1067867

Bowman, R., Cooney, O., Newbold, J. W., Thieme, A., Clark, L., Doherty, G., & Cowan, B. R. (2024). Exploring how politeness impacts the user experience of chatbots for mental health support. *International Journal of Human-Computer Studies, 184*, Article 103181. https://doi.org/10.1016/j.ijhcs.2023.103181

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (Eds.). (2000). *Embodied conversational agents*. MIT Press.

Ceha, J., Ogan, A., Walker, E., & Law, E. (2021). Can a humorous conversational agent enhance learning and engagement?. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–12). ACM. https://doi.org/10.1145/3411764.3445068.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39*(5), 752–766. https://doi.org/10.1037/0022-3514.39.5.752

Chaiken, S., Liberman, A., & Eagly, A. H. (1989). Heuristic and systematic information processing within and beyond the persuasion context. In J. S. Uleman, & J. A. Bargh (Eds.), *Unintended thought* (pp. 212–252). Guilford Press.

Chen, F. F. (2007). Sensitivity of goodness-of-fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A computational approach to politeness with application to social factors. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 250–259). Association for Computational Linguistics. https://aclanthology.org/P13-1025/.

Dubiel, M., Daronnat, S., & Leiva, L. A. (2022). Conversational agents trust calibration: A user-centred perspective to design. In *Proceedings of the 4th international conference on conversational user interfaces*. https://doi.org/10.1145/3543829.3544518 (Article 30). ACM.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Fast, E., Chen, B., & Bernstein, M. S. (2016). Empath: Understanding topic signals in text. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 4647–4657). ACM. https://doi.org/10.1145/2858036.2858535.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Gerlich, M. (2024). Societal perceptions and acceptance of virtual humans: Trust and ethics across different contexts. *Social Sciences, 13*(10), 516. https://doi.org/10.3390/socsci13100516

Gnewuch, U., Morana, S., Adam, M., & Maedche, A. (2018). Faster is not always better: Understanding the effect of dynamic response delays in human–chatbot interaction. In *Proceedings of the 26th European conference on information systems*. AIS.

Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards designing cooperative and social conversational agents for customer service. In *Proceedings of the 38th international conference on information systems*. AIS.

Gordon W. Cheung, & Roger B. Rensvold. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling, 9(2), 233–255. 10.1207/S15328007SEM0902_5.

Hair, J. F., William C. Black, Barry J. Babin, & Rolph E. Anderson. (2019). *Multivariate data analysis (8th ed.)*. Cengage Learning.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human–robot interaction. *Human Factors, 53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hutto, C. J., & Gilbert, E. (2014). 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). AAAI Press. https://doi.org/10.1609/icwsm.v8i1.14550

Jacovi, A., & Goldberg, Y. (2022). How transparency modulates trust in artificial intelligence. *Patterns, 3*(7), Article 100550. https://doi.org/10.1016/j.patter.2022.100550

John, L. Horn. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185. https://doi.org/10.1007/BF02289447

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Kumar, S., Itzhak, E., Edan, Y., Nimrod, G., Sarne-Fleischmann, V., & Tractinsky, N. (2022). Politeness in human–robot interaction: A multi-experiment study with non-humanoid robots. *International Journal of Social Robotics, 14*(8), 1805–1820. https://doi.org/10.1007/s12369-022-00911-z

Lee, K. M., & Nass, C. (2004). The multiple source effect and synthesized speech. *Communication Research, 31*(6), 820–849. https://doi.org/10.1177/0093650204267893

Meyers-Levy, J., & Tybout, A. M. (1989). Schema congruity as a basis for product evaluation. *Journal of Consumer Research, 16*(1), 39–54. https://doi.org/10.1086/209192

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th annual meeting of the Association for computational linguistics* (pp. 174–184). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1017.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Niculescu, A., & Banchs, R. E. (2019). Strategies to develop an engaging conversational agent for open-domain dialogue systems. *Computer Speech & Language, 56*, 1–21. https://doi.org/10.1016/j.csl.2018.12.001

OpenAI. (2023). GPT-4 technical report. *arXiv:2303.08774)*. https://arxiv.org/abs/2303.08774.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin. https://doi.org/10.15781/T29G6Z

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology*. Academic Press. https://doi.org/10.1016/S0065-2601(08)60214-2, 19, pp. 123–205.

Pornpitakpan, C. (2004). The persuasiveness of source credibility. *Journal of Applied Social Psychology, 34*(2), 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879–891. https://doi.org/10.3758/BRM.40.3.879

Reeves, B., & Nass, C. (1996). *The media equation*. CSLI Publications.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP–IJCNLP 2019* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410.

Richard Landis, J., & Gary, G. Koch. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., … Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics, 26*(3), 339–374. https://doi.org/10.1162/089120100561737

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger, & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73–100). MIT Press.

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. ACM. https://doi.org/10.1145/3290605.3300768.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005