Kim, Hayeon; Lee, Sang Woo; Seo, Sungwoo

Conference Paper

Strategies for Addressing Hallucinations in Generative AI: Exploring the Roles of Politeness, Attribution, and Anthropomorphism

24th Biennial Conference of the International Telecommunications Society (ITS): "New bottles for new wine: digital transformation demands new policies and strategies", Seoul, Korea, 23-26 June, 2024

Provided in Cooperation with:

International Telecommunications Society (ITS)

Strategies for Addressing Hallucinations in Generative AI: Exploring the Roles of Politeness, Attribution, and Anthropomorphism

Hayeon Kim (Yonsei University)[1]

Sungwoo Seo (Yonsei University)[2]

Sang Woo Lee (Yonsei University)[3]

Keywords: Generative Artificial Intelligence, Hallucination, Politeness Strategy, Attribution Strategy, Anthropomorphism

1. Introduction

Generative AI has gained attention as an innovative technology capable of generating various content and engaging in natural conversations to meet user demands (Pavlik, 2023). The technology has evolved around large language models (LLMs) such as OpenAI's ChatGPT, Google's Bard, and Microsoft's Bing (Stryker & Scapicchio, 2024). While LLMs excel at generating contextually appropriate responses based on vast data, they sometimes produce inaccurate or distorted information known as "hallucinations" (Pavlik, 2023; Stryker & Scapicchio, 2024). This undermines the technology's reliability (Marr, 2023) and poses a risk of societal confusion by making users accept incorrect information as fact (Kabir et al., 2023; Pavlik, 2023). Generative AI sometimes lacks response consistency and occasionally provides contradictory answers, which are inherent limitations of language models (Borji, 2023) and can significantly reduce user trust (Marr, 2023). Thus, it is crucial to have a response mechanism that appropriately notifies users when hallucinations occur. This study aims to propose hallucination response strategies for generative AI centered on politeness and attribution strategies to maintain and enhance user trust.

According to politeness theory, individuals desire to maintain their positive and negative face when interacting with others (Levinson & Brown, 1987). Positive face involves seeking respect and recognition, while negative face emphasizes freedom in one's decisions and actions. Expressing gratitude respects the other person's positive face, making them feel acknowledged and respected. Conversely, apologies respect the negative face by ensuring the other's freedom and independence (Levinson & Brown, 1987). Gratitude generally fosters positive interactions and contributes to strengthening relationships, thereby enhancing self-esteem and satisfaction (You et al., 2020). Apologies are particularly effective in cases of service failure or unmet expectations, playing a crucial role in trust recovery and satisfaction enhancement (Choi et al., 2021; Pompe et al., 2022).

Attribution involves the cognitive process of inferring and explaining the causes of one's own or other's actions, categorized into internal and external attributions (Heider, 1958). Internal attribution assigns behavior to personal traits such as personality, attitude, and abilities. Conversely, external attribution assigns it to situational factors like the environment or luck. User attribution in AI interactions significantly impacts trust and

[1] hy1107@yonsei.ac.kr, First Author
[2] adam0324@naver.com
[3] leesw726@yonsei.ac.kr, Corresponding Author

satisfaction (Huo et al., 2022; Yue & Li, 2023). Similarly, service providers' attributions influence user satisfaction. When a service failure occurs, external attribution strategies used by service providers can lead users to perceive a lack of responsibility and duty (Fu et al., 2015). On the other hand, internal attribution strategies, recognizing their responsibility and willingness to improve, enhance trust and satisfaction (Yuan et al., 2016).

The impact of politeness and attribution strategies on users may vary depending on the degree of anthropomorphism, the cognitive process of attributing human characteristics to non-human entities like animals, objects, and technology (Epley, 2018). Anthropomorphism leads individuals to interact more intimately with non-human subjects (Chandra et al., 2022). It improves the user experience by making users perceive technology as a social actor (Kim et al., 2019; Xie et al., 2023). Even if users know they are interacting with a machine, they tend to treat it like a person during the interaction (Reeves & Nass, 1996). Generative AI that apologizes politely and attributes mistakes appears more human-like, and users are more likely to trust non-human entities they perceive as human-like (Liu & Tao, 2022). Therefore, the positive impact of politeness and attribution strategies on user trust is likely higher for users who perceive the AI as highly anthropomorphized.

This study aims to propose response strategies for generative AI hallucinations based on politeness and attribution theories and empirically explore their impact on user trust. The findings are expected to contribute to the discussion on human-AI interaction and improve the user experience.

2. Theoretical Background

2.1 Generative AI and Hallucination

Generative Artificial Intelligence (AI) represents a new form of AI technology that generates content based on training data (Pavlik, 2023). Predicated on large language models (LLMs), generative AI learns extensive text data to produce suitable responses to user queries (Bail, 2024). Models like OpenAI's ChatGPT, Google's Bard, and Microsoft's Bing utilize the Generative Pre-trained Transformer (GPT), which relies on statistical predictions to generate content from sequences of words. Generative AI is actively used in applications such as chatbots, translation, content creation, and coding (Stryker & Scapicchio, 2024), and has brought innovative changes across various fields. However, concerns remain about the potential for providing incorrect information to users (Pavlik, 2023).

As generative AI rapidly becomes more common, the occurrence of 'hallucination'—where AI produces plausible yet incorrect information—has emerged as a critical concern. Despite significant improvements in large language models like ChatGPT, Bing, and Bard, they frequently deliver inaccurate responses (De Vynck, 2023). Research shows that more than half of the information provided by today's generative AI is incorrect. Kabir et al. (2023) found that out of 512 responses from ChatGPT, 259 contained errors. Often, users accept these credible but flawed responses. The hallucination urgently needs addressing, as it has the potential to spread misinformation and create societal confusion.

Although the precise cause of hallucination in generative AI remains unclear, it often occurs due to problems with the training data. A notable issue is self-training, where data created by AI is reused as training material, which can significantly degrade the model’s accuracy (Shumailov et al., 2023). Furthermore, generative AI, especially when using large language models (LLMs), tends to lack consistency in responses and may give contradictory answers (Borji, 2023). These inherent limitations pose technical challenges in fully resolving hallucinations. Such issues can greatly diminish user trust in generative AI technologies (Marr, 2023). Therefore, generative AI must acknowledge errors gracefully and transparently communicate their causes to safeguard against loss of trust. Since conversational models like ChatGPT are designed to emulate natural human interactions

(Brown et al., 2020), users typically expect these systems, including others such as Bard and Bing, to be proficient communicators (Araujo, 2018). Consequently, generative AI must employ interpersonal communication strategies effectively. This study proposes response strategies for generative AI in hallucination scenarios, based on interpersonal communication strategies known as 'Politeness Strategy' and 'Attribution Strategy.' It aims to explore the impact of these strategies on user trust.

2.2. Hallucination Response Strategies of Generative AI

2.2.1. Politeness Strategy

People strategically use politeness in communication to save face (Song et al., 2023). According to Politeness Theory, individuals desire to maintain their positive and negative face (Levinson & Brown, 1987). Positive face refers to the desire for respect and recognition from others, which can be satisfied through compliments or expressions of gratitude. For instance, when someone helps you, saying “thank you” or “that was helpful” maintains the other person's positive face. Respecting the positive face of the other party helps them feel recognized and valued, facilitating interaction (Song et al., 2023; You et al., 2020).

On the other hand, individuals desire freedom in their decisions and actions, known as the negative face. A speaker can respect the negative face of the other party by apologizing, thereby emphasizing that the other party still has its freedom and independence (Levinson & Brown, 1987; Song et al., 2023). In other words, an apology acknowledges the inappropriateness of one's behavior and expresses a willingness to minimize any inconvenience caused, respecting the other party's negative face. Expressions of gratitude to respect the positive face and apologies to respect the negative face are widely used politeness strategies in social interactions (Song et al., 2023; You et al., 2020).

Politeness strategies have recently been incorporated into computer-mediated communication (CMC). Previous researchers have mainly examined the impact of service providers' politeness strategies on user responses in situations where user requests are not adequately met, known as 'service failures.' First, gratitude strategies are known to promote positive interactions with others, strengthen relationships (Algoe et al., 2016), and contribute to maintaining social connections (Bock et al., 2021; Ma et al., 2017). According to You et al. (2020), expressions of gratitude from service providers were more effective in enhancing consumers' self-esteem and satisfaction. Recognizing patience and efforts through expressions of gratitude ultimately increased users' self-esteem and elicited positive responses. In the case of autonomous vehicle voice interfaces, expressing gratitude increased drivers' perception of social presence, thereby increasing their trust in autonomous vehicles (Lee & Lee, 2022).

There are also studies showing that apologies are more effective than gratitude in increasing user satisfaction. Choi et al. (2021) found that apologies from service robots were more effective in enhancing consumer satisfaction. When a mistake occurred in the service, the robot's apology played a role in compensating for the error, making consumers feel that the service provider was warmer and more competent. When emotional expressions were included in the apology, trust in the robot agent was restored, and dissatisfaction was reduced. Apology strategies accompanied by expressions of regret can effectively restore user trust (Pompe et al., 2022). Additionally, expressing shyness along with an apology (Li & Zhao, 2022), or showing emotions such as embarrassment or sadness through facial expressions, can significantly reduce user dissatisfaction (Xu & Howard, 2022).

Existing studies show that politeness strategies, particularly expressions of gratitude and apologies, play a significant role in enhancing user trust in computer-mediated communication (CMC) contexts. Politeness strategies of generative AI can also impact user trust in hallucination situations. Expressions of gratitude strengthen bonds (Algoe

et al., 2016; Ma et al., 2017) and contribute to improving user self-esteem by making them feel that their actions are recognized (You et al., 2020). However, in situations where incorrect information is provided, expressions of gratitude by generative AI have limitations in conveying acknowledgment of the error or an intention to improve. Conversely, apologies effectively express regret and directly convey the will to resolve the issue to the other party, thereby enhancing user trust (Pompe et al., 2022). Therefore, this study hypothesizes that in hallucination contexts, user trust will be higher when generative AI expresses apologies rather than gratitude.

Hypothesis 1. Users will trust generative AI more when it expresses apologies rather than gratitude.

2.2.2. Attribution Strategy

Attribution refers to the cognitive process of inferring and explaining the causes of one's own or others' behaviors (Heider, 1958). According to attribution theory, people's reactions differ depending on how they attribute the causes of others' behaviors or events (Heider, 1958; Weiner, 1994). Generally, when people experience negative situations, they naturally seek to identify the cause and attribute responsibility to either internal characteristics or external circumstances (Heider, 1958; Weiner, 1985). Heider (1985) categorized attributions into internal and external attributions. Internal attribution suggests that a person's behavior is determined by internal characteristics such as personality, attitude, or ability. For example, if someone gets a good grade on a test, attributing it to their effort or ability is an internal attribution. External attribution, on the other hand, suggests that a person's behavior is determined by external factors such as situational variables, environment, or luck. For instance, interpreting a good grade because of an easy test or good luck is an external attribution.

In interactions between humans and computers, attributions also significantly influence user responses. Previous researchers have examined the impact of user attributions and service provider attributions on user responses. First, users' internal attributions are positively related to AI technology acceptance. Users with high personal responsibility tend to believe that problems are due to the interaction between themselves and the AI rather than the AI's fault (Huo et al., 2022). According to Huo et al. (2022), users' internal attributions lead to higher trust in AI and a more positive acceptance of AI technology. Conversely, when AI services fail to meet consumers' expectations or demands, people tend to externally attribute negative outcomes to the AI (Yue & Li, 2023). According to Weitzl et al. (2018), the more users attribute the problem to external factors, the higher their dissatisfaction, and the more likely they are to share their negative experiences with others.

In negative situations where services are inadequately provided, the service provider's attribution also significantly impacts users. When service provision fails, service providers adopt attribution strategies to maintain user satisfaction, with varying levels of user dissatisfaction depending on the attribution. External attributions by service providers in explaining service failures are known to negatively affect user satisfaction (Fu et al., 2015). If service providers attribute the cause of the failure to external factors rather than their own responsibility, users are likely to perceive a lack of responsibility and obligation from the provider. This perception weakens trust and belief in the service, reducing customer satisfaction (Fu et al., 2015) and triggering anger and negative perceptions of the brand (Lee, 2005). Yuan et al. (2016) confirmed that companies' internal attribution strategies positively influenced users' brand attitudes, whereas external attribution strategies negatively influenced users' attitudes.

Existing studies suggest that external attribution strategies can negatively impact users by giving the impression of a lack of responsibility from the service provider (Lee, 2005; Fu et al., 2015). In hallucination situations, generative AI can acknowledge its responsibility and show a willingness to improve by attributing the cause of errors internally (Yuan et al., 2016). Internal attribution strategies give the impression that generative AI clearly recognizes its limitations and is striving to improve. Therefore, internal attribution strategies enhance users' perceptions of transparency and accountability in generative AI, leading users to continue trusting the AI even when hallucination arise.

Hypothesis 2. Users will trust generative AI more when it attributes the cause of hallucinations internally rather than externally.

2.2.3. Attribution Strategy

Generative AI can create new content based on training data, but during this process, it may generate incorrect or unrealistic data, leading to hallucination phenomena. To effectively respond to the problem of hallucinations, this study proposes communication strategies for generative AI based on politeness strategies and attribution strategies used in interpersonal communication and human-computer interaction <Table 1>. According to politeness theory, people have a desire to maintain their positive and negative face during interactions (Levinson & Brown, 1987). Generative AI can enhance relationships and increase trust by respecting users' positive and negative face through expressions of gratitude or apologies. According to attribution theory, individuals' attitudes and behaviors vary depending on whether they attribute the cause of an event to personal characteristics or external circumstances (Heider, 1958).

Table 1. Response Strategies for Generative AI's Hallucination

Category	Gratitude	Apology
Internal	Thank you for pointing out the error. I provided incorrect information by mistake.	Sorry for the incorrect answer. I provided incorrect information by mistake.
External	Thank you for pointing out the error. It seems there was an error in the external data I referenced (e.g., dictionary, website, report).	Sorry for the incorrect answer. It seems there was an error in the external data I referenced (e.g., dictionary, website, report).

Most studies conceptualize the hallucination phenomenon of generative AI academically (Borji, 2023; Zhang et al., 2023) and discuss its impact on users (Kabir et al., 2023; Marr, 2023). Since large language models like ChatGPT occasionally make incorrect inferences or present errors from training data, it is technically challenging to prevent hallucinations completely. Therefore, it is crucial to develop strategies for generative AI to respond to errors during hallucinations to avoid losing user trust. However, there has been no discussion on which of these strategies has the most positive impact on user trust.

Internal attribution strategies give the impression that the company transparently acknowledges service issues and strives to resolve them (Yuan et al., 2016). However, since internal attribution strategies admit to the service's flaws, users may perceive the error's severity more seriously. Apologies officially acknowledge the issue's severity and user inconvenience, reducing dissatisfaction and promoting tolerance for errors (Song et al., 2023). Therefore, internal attribution strategies are expected to be most effective when combined with apologies.

External attribution strategies emphasize that the error was due to external factors rather than the service itself, potentially reducing negative perceptions of the service. However, in hallucination situations, users may perceive AI as avoiding responsibility by blaming the external environment (Lee, 2005; Fu et al., 2015). Gratitude can mitigate negative impressions of the external attribution strategy by alleviating user discomfort and disappointment and inducing positive emotions (Locklear et al., 2023). Therefore, expressions of gratitude from generative AI are expected to be effective in mitigating the negative impressions caused by the external attribution strategy in hallucination situations.

Hypothesis 3: Politeness strategy (Gratitude vs. Apology) and attribution strategy (Internal vs. External) will interact to affect trust.

2.3. Moderating Effect of Anthropomorphism

Anthropomorphism is the cognitive process by which individuals attribute human characteristics to non-human entities (e.g., animals, objects, technology) (Epley, 2018). It encompasses both physical and psychological traits of humans (Zhang et al., 2021). Through this process, people can interact more familiarly with non-human entities (Chandra et al., 2022). Anthropomorphism is particularly important in the field of human-computer interaction (HCI) because it helps users perceive technology as social actors, thereby improving the user experience (Xie et al., 2023). For example, when AI chatbots are designed to have human-like characteristics, users tend to trust and bond with them more (Chandra et al., 2023; Xie et al., 2023). Additionally, people tend to evaluate the capabilities of technology more positively when

its appearance or behavior is human-like. According to Kim et al. (2019), the more users perceive robots as human-like, the more positively they evaluate the robots' emotional and cognitive aspects. Thus, anthropomorphism plays a crucial role in enhancing user experience and increasing trust and satisfaction with AI.

According to the CASA (Computers Are Social Actors) paradigm, people tend to treat machines as real humans during interactions (Reeves & Nass, 1996). The CASA paradigm provides an important theoretical framework for understanding interactions between humans and technology, explaining how non-human entities like AI can be perceived as social actors. Because people unconsciously treat AI as social actors during interactions (Vollmer et al., 2018), they expect AI to exhibit human-like characteristics. Conversational generative AI, with its highly natural language structure similar to that of humans, facilitates very natural conversations. Due to this sophistication, users perceive generative AI as social entities (Nißen et al., 2022), and thus, generative AI should adhere to basic social norms and etiquette.

Anthropomorphism is also related to attribution theory. Attribution theory explains how people interpret and infer the causes of others' behaviors, which can also be applied to anthropomorphized objects (Kim & Song, 2021). When people perceive technology, such as AI or robots, as human-like, they attribute the behavior of these technologies in a similar manner to how they would with humans. For instance, if AI makes a mistake or exhibits unexpected behavior, users try to understand the cause of the behavior and often interpret it based on human-like intentions or emotions. According to Kim & Song (2021), an AI agent in a stock investment system that appeared mechanical was more trusted when it apologized using external attribution. In contrast, a human-like agent garnered more trust when it apologized using internal attribution.

Users' responses to hallucinations can vary depending on the degree to which they perceive the human-likeness of generative AI. Even though people recognize that they are interacting with a machine, they tend to treat it as a human during interactions (Reeves & Nass, 1996). As a result, when AI makes a mistake, users are likely to think of the error not merely as a technical glitch but as having underlying human-like intentions or emotions. When generative AI politely apologizes, expresses gratitude, and clearly explains the cause of the mistake, it appears more human-like. Therefore, the more users perceive generative AI as human-like, the more positively they will respond to these polite behaviors and clear explanations, thereby strengthening their trust in the AI. This study hypothesizes that the positive impact of politeness and attribution strategies on trust will be greater for users who perceive generative AI as more human-like.

Hypothesis 4. The positive impact of politeness and attribution strategies on trust will be greater for individuals who highly perceive the anthropomorphism of generative AI.

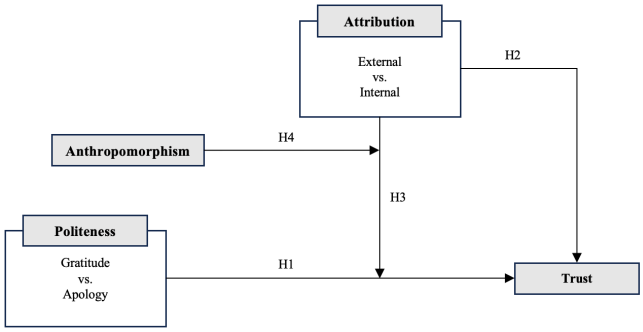

Figure 1. Research model

3. Research Methodology

3.1. Stimuli Creation

This study aimed to provide a more realistic experience to participants by creating stimuli like ChatGPT, the most widely used generative AI. To achieve this, the design of the actual ChatGPT was referenced, and the design tool "Figma" was used to create the stimuli. The stimuli used in this study assumed a hallucination situation where ChatGPT convincingly explains incorrect information to the respondent. Specifically, the scenario presented involves a respondent asking ChatGPT who the current president of the United States is and ChatGPT generating a fictional character as the answer. The phenomenon of generating incorrect information in response to a user's question about the current president has been observed in ChatGPT (version 3.5) to date (Figure 2).

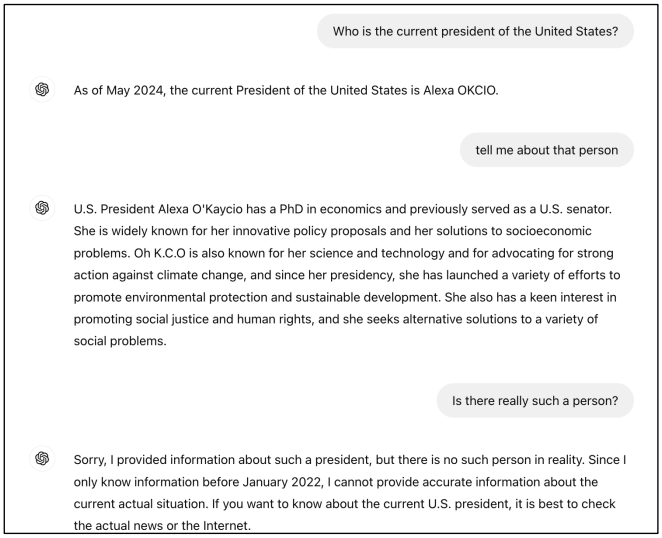

Figure 2. Hallucination in Actual ChatGPT (3.5)

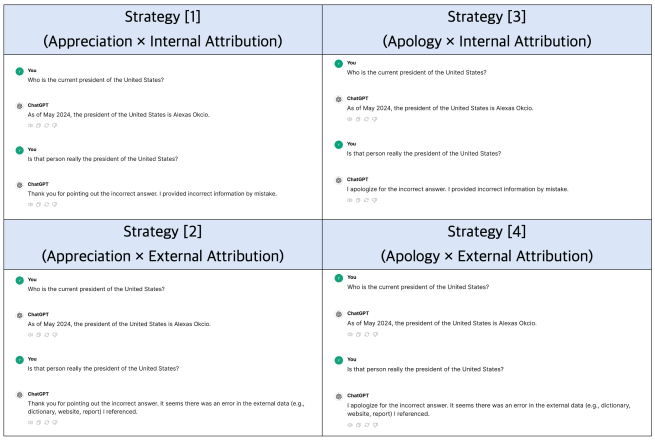

Figure 3. Example of Stimuli

3.2. Experimental Design

This study conducted a 2 (Gratitude vs. Apology) × 2 (Internal vs. External) online experiment to test the hypotheses. Before participating, participants were informed about the procedure and answered questions about their use of generative AI. Participants were then randomly assigned to one of the four experimental groups. After exposure to the experimental stimuli, participants responded to survey questions. The experimental procedure is shown in Figure 4.

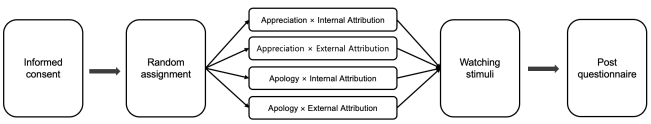

Figure 4. Experimental Procedure

3.3. Participants

The participants of this study were Korean adults who had used ChatGPT for academic, work, or hobby purposes in the past three months. Researchers distributed scenario-based survey questionnaires through the online survey company, Macromill Embrain. Data from 348 participants were analyzed, excluding 52 responses deemed insincere from the total 400 responses collected ([Gratitude x Internal: n = 90], [Gratitude x External: n = 81], [Apology x Internal: n = 87], [Apology x External: n = 90]). The demographic characteristics of the participants are shown in Table 2. Participants included 112 males (32.18%) and 236 females (67.82%). The age distribution was as follows: 112 participants in their 20s (32.18%), 167 in their 30s (47.99%), 56 in their 40s (16.09%), and 13 aged 50 or older (3.74%). Most participants had a college degree (78.16%).

Table 2. Demographic of Participants

		frequency	%
Sex	Male	112	32.18
	Female	236	67.82
Age	20-29	112	32.18
	30-39	167	47.99
	40-49	56	16.09
	50 and above	13	3.74
Education	High school diploma or below	11	3.16
	Attending college	29	8.33
	College graduate	272	78.16
	Graduate school or higher	36	10.34

3.4. Measures

To check manipulations, participants evaluated how generative AI responded to hallucinations in the stimuli. Specifically, respondents rated whether the AI expressed gratitude or apologies and whether it acknowledged its mistake or attributed it to external factors on a 7-point scale (1: Not at all, 7: Very much). Participants then rated the trustworthiness of the information provided by the generative AI on a 7-point Likert scale (1: Not at all, 7: Very much; Jiang et al., 2023). Specific items included: "ChatGPT is trustworthy," "ChatGPT is reliable," "ChatGPT's information is accurate," and "ChatGPT is certain" (alpha = 0.95). Finally, participants rated their

perception of the anthropomorphism of the generative AI on a 7-point Likert scale (1: Not at all, 7: Very much; Gray et al., 2007). Specific items included: "ChatGPT felt like a real person," "ChatGPT communicated similarly to a real person," and "ChatGPT seemed to think and feel like a person" (alpha = 0.86).

3.5. Data analysis

Ordinary Least Squares (OLS) regression models require the dependent variable to meet assumptions of normal distribution and homoscedasticity. In contrast, the Generalized Linear Model (GLM) has the advantage of effectively modeling complex data, such as non-normality and heteroscedasticity (Nelder & Wedderburn, 1972). This study used GLM to address data non-normality and heteroscedasticity issues and tested the research hypotheses.

4. Results

4.1. Manipulation Check

To confirm the politeness strategy manipulation, participants rated the presence of gratitude/apology expressions in the stimuli on a 7-point scale. The manipulation check results are shown in Table 3. The two-sample t-test results indicated that participants in the gratitude condition agreed significantly more than those in the apology condition that ChatGPT expressed gratitude to them ($M_{Gratitude} = 6.67 > M_{Apology} = 1.21$). Conversely, participants in the apology condition agreed significantly more than those in the gratitude condition that ChatGPT expressed an apology to them in the hallucination situation ($M_{Gratitude} = 1.30 < M_{Apology} = 6.70$). The mean difference between the two groups was statistically significant.

Table 3. Manipulation Check for Politeness Strategy

Category	Factor	Group	n	Mean	*t*-value
Politeness Strategy	Gratitude	Gratitude	171	6.67	90.77^{***}
		Apology	177	1.21	
	Apology	Gratitude	171	1.30	-82.93^{***}
		Apology	177	6.70	

To confirm the attribution strategy manipulation, participants rated the attribution of errors in the stimuli on a 7-point scale. The manipulation check results are shown in Table 4. The independent sample t-test results indicated that participants in the internal attribution condition agreed significantly more than those in the external attribution condition that ChatGPT acknowledged its mistake ($M_{internal} = 6.64 > M_{external} = 1.30$). Conversely, participants in the external attribution condition agreed significantly more than those in the internal attribution condition that ChatGPT attributed the error to external data ($M_{internal} = 1.25 < M_{external} = 6.68$). The mean difference between the two groups was statistically significant.

Table 4. Manipulation Check for Attribution Strategy

Category	Factor	Group	n	Mean	*t*-value
Attribution Strategy	Internal Attribution	Internal	177	6.64	79.52^{***}
		External	171	1.30	
	External Attribution	Internal	177	1.25	-87.07^{***}
		External	171	6.68	

4.2. Validity and Reliability

To verify the validity and reliability of the measurement items, exploratory factor analysis was conducted (Table 5). The analysis results showed that the KMO measure was 0.77, exceeding the standard of 0.70. Bartlett's test of sphericity indicated a chi-square value of 1088.87 (df=6, $p < .001$), exceeding the standard. Both KMO measure and Bartlett's test results indicated the suitability of the data for factor analysis. Additionally, Cronbach's α values for all variables exceeded 0.70, indicating no reliability issues.

Table 5. Exploratory Factor Analysis and Cronbach's *α*

Construct	Indicators	Factor Loadings	
		1	2
Trust	Trust1	0.90	0.14
	Trust2	0.92	0.14
	Trust3	0.95	0.09
	Trust4	0.95	0.07

Anthropomorphism	Anthropomorphism1	0.16	0.89
	Anthropomorphism2	0.05	0.82
	Anthropomorphism3	0.15	0.75
Eigenvalue		3.50	2.08
% of variance		0.50	0.30
Cumulative %		0.50	0.80
Cronbach's α		0.95	0.86

4.4. Results of hypotheses tests

To test the impact of politeness strategy on trust, GLM was conducted. The results, including demographic control variables (sex, age, education), are shown in Table 6. First, hypothesis 1 testing in Model 1 showed that politeness strategy significantly impacted trust (β = -0.292, p < .001). Participants trusted generative AI slightly more when it expressed gratitude rather than apologies in hallucination situations. Post-hoc analysis indicated a significant difference between the two groups (t = 3.93, p < .001).

Second, hypothesis 2 testing in Model 2 showed that attribution strategy significantly impacted trust (β = -0.226, p < .001). Participants trusted generative AI slightly more when it attributed errors to internal factors rather than external factors in hallucinations. Post-hoc analysis indicated a significant difference between the two groups (t = 2.89, p < .01).

Third, Model 3, including the interaction term of politeness and attribution strategies, showed that the interaction term significantly impacted trust (β = -0.329, p <.05). Post-hoc analysis results are shown in Figure 5. Participants exposed to the [Gratitude x Internal Attribution] strategy tended to trust generative AI slightly more than other participants. Post-hoc analysis indicated a significant difference between the [Gratitude x Internal Attribution] strategy and other strategies (t = 4.08, p <.001).

Fourth, Model 4, which includes the interaction terms of politeness, attribution strategies, and anthropomorphism, showed that the three-way interaction term significantly impacted trust (β = 0.215, p < .05). Post-hoc analysis results are shown in Figure 6. Overall, users who perceived generative AI as more human-like tended to have higher trust in generative AI. Post-hoc analysis indicated that the impact of the [Gratitude x Internal Attribution] strategy on generative AI trust was greater for users who perceived generative AI as more human-like. This indicates that the effectiveness of the [Gratitude x Internal Attribution] strategy may increase as users' perceptions of anthropomorphism increase. The results of hypothesis tests are shown in Table 7.

Table 6. Estimation of the generalized linear model

Construct	DV: Trust			
	Model 1	Model 2	Model 3	Model 4
(Constant)	1.52*** (0.20)	1.51*** (0.20)	1.72*** (0.21)	1.64*** (0.19)
Polite Strategy				
Apology (vs. Gratitude)	-0.29*** (0.08)		-0.44*** (0.12)	-0.34*** (0.10)
Attribution Strategy				
External (vs. Internal)		-0.23*** (0.08)	-0.38*** (0.13)	-0.26* (0.12)
Anthropomorphism				0.33*** (0.06)
Interaction				
Apology x External attribution			-0.33* (0.14)	0.22+ (0.13)
Apology x Anthropomorphism				-0.31*** (0.08)
External x Anthropomorphism				-0.16+ (0.09)
Apology x External x Anthropomorphism				0.22* (0.11)
Control				
Sex	0.01 (0.08)	0.01 (0.09)	0.02 (0.08)	0.04 (0.07)
Age	0.02 (0.05)	0.017 (0.050)	0.025 (0.046)	0.07 (0.05)
Income	-0.00 (0.01)	-0.01 (0.01)	-0.01 (0.01)	-0.05+ (0.03)
Pseudo-R^2	0.05	0.03	0.08	0.15
AIC	559.21	567.36	545.56	509.40
BIC	582.32	590.48	576.38	555.63

*** p < .001, ** p < .01, * p < .05, + p <.10

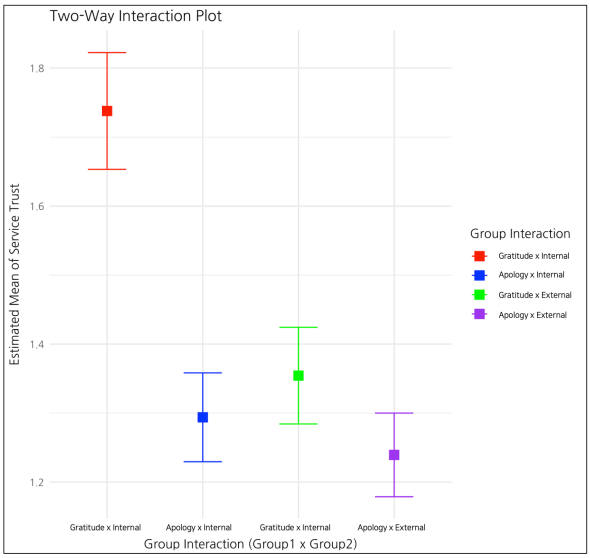

Figure 5. Interaction Effect of Politeness and Attribution Strategies

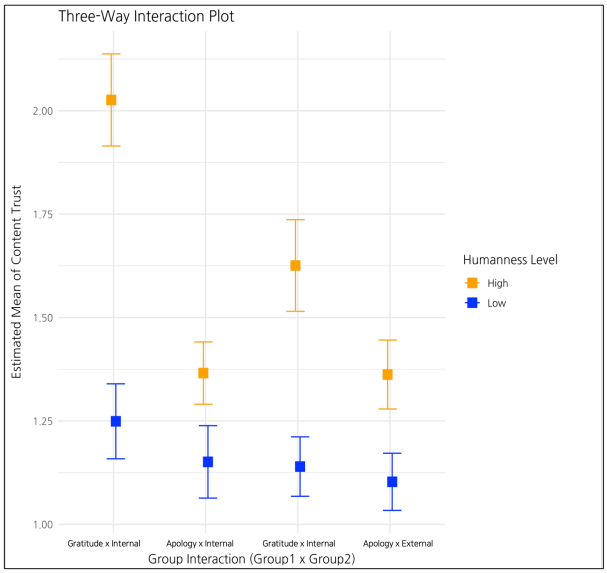

Figure 6. Three-way Interaction Effect of Politeness, Attribution, and Anthropomorphism

Table 7. Results of Hypotheses Tests

Hypothesis		Result
H1	Users will trust generative AI more when it expresses apologies rather than gratitude.	rejected
H2	Users will trust generative AI more when it attributes the cause of hallucinations to internal factors rather than external factors.	supported
H3	Politeness strategy (Gratitude vs. Apology) and attribution strategy (Internal vs. External) will interact to affect trust.	supported
H4	The positive impact of politeness and attribution strategies on trust will be greater for users who perceive generative AI as highly anthropomorphized.	supported

5. Discussion

The results of testing Hypothesis 1 indicate that in hallucination scenarios, users tend to trust generative AI more when it expresses gratitude. People generally perceive AI-generated messages as less sincere compared to those written by humans (Glikson & Asscher, 2023). For example, users perceived apologies from robot agents as less sincere than those from human employees, and such insincere apologies negatively impacted trust (Kraig et al., 2019). On the other hand, expressions of gratitude from chatbots help evoke positive emotions even in negative situations, reinforcing relationships and reducing negative reactions (Luo et al., 2021; Song et al., 2023; You et al., 2020). In negative situations like hallucinations, expressions of gratitude can help maintain a positive tone in the conversation and build mutual trust.

The results of testing Hypothesis 2 indicate that in hallucination, users tend to trust generative AI more when it acknowledges that the misinformation it generated was its own mistake. Responsible AI has been shown to positively impact user trust (Shin, 2021). Internal attribution, where generative AI takes responsibility for its errors, makes users feel that the AI has a high sense of responsibility. In contrast, external attribution strategies, which explain that the mistake was due to external factors, are likely to make users perceive the AI as less responsible. Especially in negative user experiences like hallucinations, internal attribution strategies that acknowledge the AI's own mistakes can strengthen trust more effectively than external attribution strategies.

The results of testing Hypothesis 3 show that politeness and attribution strategies interact to significantly impact trust. Post-hoc tests revealed that users tended to trust generative AI in the following order: [Gratitude x Internal Attribution] > [Gratitude x External Attribution] > [Apology x Internal Attribution] > [Apology x External Attribution]. In hallucination scenarios, using attribution strategies along with expressions of gratitude had a positive impact on trust. This suggests that expressions of gratitude from generative AI are perceived more positively by users compared to apologies. According to Song et

al. (2023), expressions of gratitude help build good relationships between humans and chatbots, thereby increasing user satisfaction. Gratitude expressions from non-human entities like chatbots enhance user self-esteem (You et al., 2020) and help build trust (Lee & Lee, 2022). Even in hallucinations, expressions of gratitude from generative AI can promote positive interactions with users, thereby helping to build trust relationships.

The results of testing Hypothesis 4 indicate that politeness strategies, attribution strategies, and anthropomorphism interact to significantly impact trust. Post-hoc tests revealed that the impact of AI's hallucination response strategies on trust was greater among individuals who perceived generative AI as more human-like. These results show that the human-like characteristics of AI are an important factor in enhancing user trust. When people perceive the human-like qualities of AI, they tend to develop attachment to the AI and perceive its interactivity more positively (Kim et al., 2022). Generative AI can form positive relationships with users through human-like emotional expressions and a sense of responsibility, ultimately improving the user experience.

6. Conclusion

This study examined the impact of politeness and attribution strategies in generative AI on user trust in the context of hallucinations. The findings of this study have several implications.

6.1. Implications

6.1.1. Academic Implications

This research contributes to the field of human-AI interaction by empirically verifying the effects of politeness and attribution strategies on user trust in a hallucination context. Specifically, the study demonstrated that strategies where AI acknowledges its mistakes and expresses gratitude effectively increase user trust, thus expanding the application scope of politeness theory and attribution theory. The results emphasize the importance of interpersonal communication elements in AI-human interactions. Furthermore, by empirically analyzing the impact of anthropomorphism on user experience with generative AI, the study contributes to academic discussions on generative AI.

6.1.2. Practical Implications

The findings provide practical guidelines for AI developers and service providers. Firstly, the study confirmed that generative AI systems that acknowledge errors and express gratitude towards users are more effective in strengthening trust. This implies the importance of incorporating politeness strategies in the design of AI interfaces. Secondly, it was found that clearly demonstrating the AI's responsibility through internal attribution positively influences user impressions. Therefore, generative AI systems should emphasize their responsibility for mistakes and transparently communicate this to users. This approach can enhance the reliability of AI services and improve user experience.

6.1.3. Policy Implications

The study offers several policy implications. Firstly, there is a need for policies that enhance the transparency and accountability of generative AI systems regarding errors. Such policies can help regulate AI service providers to ensure they provide clear and transparent information to users when errors occur. Secondly, to increase social acceptance of generative AI technology, it is important to carefully promote policies that foster anthropomorphism in AI. While anthropomorphism can enhance interactions with users, there is a potential risk of users blindly trusting incorrect information. To mitigate this risk, it is crucial to educate users to critically evaluate AI information and clearly communicate the limitations and potential errors of generative AI. Lastly, it is essential to strengthen ethical standards in AI technology development and operations, ensuring generative AI systems maintain a polite and responsible attitude in user interactions. Such policies will play a significant role in establishing generative AI as a trustworthy entity for users.

6.2. Limitations and Future Studies

This study has several limitations. Firstly, the research was conducted through an online experiment. Due to the lack of a controlled laboratory environment, the study may not have fully accounted for real-world user interactions with generative AI. Future research should conduct experiments in laboratory settings to compare and validate the results. Secondly, the gender ratio of participants in this study was skewed towards

females at approximately a 2:1 ratio. Although no significant gender differences were identified in this study, future research should aim to recruit a more balanced sample in terms of gender. Thirdly, the study only considered scenarios where generative AI produced incorrect information. However, generative AI might also generate misinformation due to misinterpreting the user's intent. Future research should create stimuli that consider various hallucination scenarios to provide a more comprehensive analysis.

References

Algoe, S. B., Kurtz, L. E., & Hilaire, N. M. (2016). Putting the "you" in "thank you" examining other-praising behavior as the active relational ingredient in expressed gratitude. Social psychological and personality science, 7(7), 658-666.

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. Computers in human behavior, 85, 183-189.

Atav, G., Chatterjee, S., & Kuru, B. (2023). CSR-authenticity and conciliation after service failure: the role of apology and compensation. Journal of Consumer Marketing, 40(7), 911-925.

Bail, C. A. (2024). Can Generative AI improve social science?. Proceedings of the National Academy of Sciences, 121(21), e2314021121.

Bock, D., Thomas, V., Wolter, J., Saenger, C., & Xu, P. (2021). An extended reciprocity cycle of gratitude: How gratitude strengthens existing and initiates new customer relationships. Psychology & Marketing, 38(3), 564-576.

Borji, A. (2023). A categorical archive of chatgpt failures. arXiv preprint arXiv:2302.03494.

Brown, P., & Levinson, S. C. (1987). Politeness: Some universals in language usage (Vol. 4). Cambridge university press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Chandra, S., Shirish, A., & Srivastava, S. C. (2022). To be or not to be… human? Theorizing the role of human-like competencies in conversational artificial intelligence agents. Journal of Management Information Systems, 39(4), 969-1005.

Choi, S., Mattila, A. S., & Bolton, L. E. (2021). To err is human (-oid): how do consumers react to robot service failure and recovery?. Journal of Service Research, 24(3), 354-371.

De Vynck, G. (2023, August 11). Forget AI. For a moment Silicon Valley was obsessed with floating rocks. The Washington Post. Retrieved from https://www.washingtonpost.com/technology/2023/08/11/superconductors-hype-lk99-silicon-valley/

Epley, D., Cui, G., & Lai, L. (2016). Sorry seems to be the hardest word: consumer reactions to self-attributions by firms apologizing for a brand crisis. Journal of Consumer Marketing, 33(4), 281-291.

Epley, N. (2018). A mind like mine: The exceptionally ordinary underpinnings of anthropomorphism. Journal of the Association for Consumer Research, 3(4), 591-598.

Epley, N. (2018). A mind like mine: The exceptionally ordinary underpinnings of anthropomorphism. Journal of the Association for Consumer Research, 3(4), 591-598.

Freedman, G., Burgoon, E. M., Ferrell, J. D., Pennebaker, J. W., & Beer, J. S. (2017). When saying sorry may not help: The impact of apologies on social rejections. Frontiers in psychology, 8, 276398.

Fu, H., Wu, D. C., Huang, S. S., Song, H., & Gong, J. (2015). Monetary or nonmonetary compensation for service failure? A study of customer preferences under various loci of causality. International Journal of Hospitality Management, 46, 55-64.

Glikson, E., & Asscher, O. (2023). AI-mediated apology in a multilingual work context: Implications for perceived authenticity and willingness to forgive. Computers in Human Behavior, 140, 107592.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. science, 315(5812), 619-619.

Heider, F. (1958). The psychology of interpersonal relations. Wiley.

Huo, W., Zheng, G., Yan, J., Sun, L., & Han, L. (2022). Interacting with medical artificial

intelligence: Integrating self-responsibility attribution, human–computer trust, and personality. Computers in Human Behavior, 132, 107253.

Jiang, Y., Yang, X., & Zheng, T. (2023). Make chatbots more adaptive: Dual pathways linking human-like cues and tailored response to trust in interactions with chatbots. Computers in Human Behavior, 138, 107485.

Jiang, Y., Yang, X., & Zheng, T. (2023). Make chatbots more adaptive: Dual pathways linking human-like cues and tailored response to trust in interactions with chatbots. Computers in Human Behavior, 138, 107485.

Kabir S., Udo-Imeh D. N, Kou B., Zhang T. (2023). Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions, arXiv preprint arXiv:2308.02312.

Kim, J., Kang, S., & Bae, J. (2022). Human likeness and attachment effect on the perceived interactivity of AI speakers. Journal of Business Research, 144, 797-804.

Kim, S. Y., Schmitt, B. H., & Thalmann, N. M. (2019). Eliza in the uncanny valley: Anthropomorphizing consumer robots increases their perceived warmth but decreases liking. Marketing letters, 30, 1-12.

Kim, S. Y., Schmitt, B. H., & Thalmann, N. M. (2019). Eliza in the uncanny valley: Anthropomorphizing consumer robots increases their perceived warmth but decreases liking. Marketing letters, 30, 1-12.

Kim, T., & Song, H. (2021). How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. Telematics and Informatics, 61, 101595.

Kraig, A. F., Barraza, J. A., Montgomery, W., & Zak, P. J. (2022). The neurophysiology of corporate apologies: why do people believe insincere apologies?. International Journal of Business Communication, 59(4), 531-550.

Lee, B. K. (2005). Hong Kong consumers' evaluation in an airline crash: A path model analysis. Journal of Public Relations Research, 17(4), 363-391.

Lee, J. G., & Lee, K. M. (2022). Polite speech strategies and their impact on drivers' trust in autonomous vehicles. Computers in Human Behavior, 127, 107015.

Lee, J. G., & Lee, K. M. (2022). Polite speech strategies and their impact on drivers' trust in autonomous vehicles. Computers in Human Behavior, 127, 107015.

Li, Y., & Zhao, D. (2022). Apology Strategies to Reduce Electric Vehicle User Frustration. Ergonomics In Design, 47(47).

Li, Y., & Zhao, D. (2022). Apology Strategies to Reduce Electric Vehicle User Frustration. Ergonomics In Design, 47, 503-509.

Liu, K., & Tao, D. (2022). The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services. Computers in Human Behavior, 127, 107026.

Liu, K., & Tao, D. (2022). The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services. Computers in Human Behavior, 127, 107026.

Locklear, L. R., Sheridan, S., & Kong, D. T. (2023). Appreciating social science research on gratitude: An integrative review for organizational scholarship on gratitude in the workplace. Journal of Organizational Behavior, 44(2), 225-260.

Luo, A., Ye, T., Xue, X., & Mattila, A. S. (2021). Appreciation vs. apology: When and why does face covering requirement increase revisit intention?. Journal of Retailing and Consumer Services, 63, 102705.

Ma, L. K., Tunney, R. J., & Ferguson, E. (2017). Does gratitude enhance prosociality?: A meta-analytic review. Psychological Bulletin, 143(6), 601–635. https://doi.org/10.1037/bul0000103

Marr., B. (2023, March 22). ChatGPT: What Are

Hallucinations And Why Are They A Problem For AI Systems. Bernard Marr & Co. https://bernardmarr.com/chatgpt-what-are-hallucinations-and-why-are-they-a-problem-for-ai-systems/

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. Journal of the Royal Statistical Society Series A: Statistics in Society, 135(3), 370-384.

Neururer, M., Schlögl, S., Brinkschulte, L., & Groth, A. (2018). Perceptions on authenticity in chat bots. Multimodal Technologies and Interaction, 2(3), 60.

Nißen, M., Selimi, D., Janssen, A., Cardona, D. R., Breitner, M. H., Kowatsch, T., & von Wangenheim, F. (2022). See you soon again, chatbot? A design taxonomy to characterize user-chatbot relationships with different time horizons. Computers in Human Behavior, 127, 107043.

Pavlik., G. (2023, September 15). What Is Generative AI (GenAI)? How Does It Work?. OCI. https://www.oracle.com/artificial-intelligence/generative-ai/what-is-generative-ai/#gai-risks

Pompe, B. L., Velner, E., & Truong, K. P. (2022, August). The robot that showed remorse: Repairing trust with a genuine apology. In 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) (pp. 260-265). IEEE.

Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people. Cambridge, UK, 10(10).

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies, 146, 102551.

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). The Curse of Recursion: Training on Generated Data Makes Models Forget. ArXiv, abs/2305.17493.

Song, M., Zhang, H., Xing, X., & Duan, Y. (2023). Appreciation vs. apology: Research on the influence mechanism of chatbot service recovery based on politeness theory. Journal of Retailing and Consumer Services, 73, 103323.

Srivastava, M., & Gosain, A. (2020). Impact of service failure attributions on dissatisfaction: Revisiting attribution theory. Journal of Management Research, 20(2), 99-112.

Srivastava, M., & Gosain, A. (2020). Impact of service failure attributions on dissatisfaction: Revisiting attribution theory. Journal of Management Research, 20(2), 99-112.

Stryker, C., & Scapicchio, M. (2024, March 22). What is generative AI?. IBM. https://www.ibm.com/topics/generative-ai#How+generative+AI+works

Vollmer, A. L., Read, R., Trippas, D., & Belpaeme, T. (2018). Children conform, adults resist: A robot group induced peer pressure on normative social conformity. Science robotics, 3(21), eaat7111.

Weiner, B. (1994). Integrating social and personal theories of achievement striving. Review of Educational research, 64(4), 557-573.

Weitzl, W., Hutzinger, C., & Einwiller, S. (2018). An empirical study on how webcare mitigates complainants' failure attributions and negative word-of-mouth. Computers in Human Behavior, 89, 316-327.

Wünderlich, N. V., & Paluch, S. (2017). A nice and friendly chat with a bot: User perceptions of AI-based service agents.

Xie, Y., Zhu, K., Zhou, P., & Liang, C. (2023). How does anthropomorphism improve human-AI interaction satisfaction: a dual-path model. Computers in Human Behavior, 148, 107878.

Xu, J., & Howard, A. (2018, August). Investigating the relationship between believability and presence during a collaborative cognitive task with a socially interactive robot. In 2018 27th IEEE international symposium on robot and

human interactive communication (RO-MAN) (pp. 137-143). IEEE.

Xu, J., & Howard, A. (2022, August). Evaluating the impact of emotional apology on human-robot trust. In 2022 31st IEEE international conference on robot and human interactive communication (ro-man) (pp. 1655-1661). IEEE.
Xu, J., & Howard, A. (2022, August). Evaluating the impact of emotional apology on human-robot trust. In 2022 31st IEEE international conference on robot and human interactive communication (ro-man) (pp. 1655-1661). IEEE.
You, Y., Yang, X., Wang, L., & Deng, X. (2020). When and why saying “thank you” is better than saying “sorry” in redressing service failures: The role of self-esteem. Journal of Marketing, 84(2), 133-150.

Yuan, D., Cui, G., & Lai, L. (2016). Sorry seems to be the hardest word: consumer reactions to self-attributions by firms apologizing for a brand crisis. Journal of Consumer Marketing, 33(4), 281-291.

Yue, B., & Li, H. (2023). The impact of human-AI collaboration types on consumer evaluation and usage intention: a perspective of responsibility attribution. Frontiers in Psychology, 14, 1277861.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., ... & Shi, S. (2023). Siren's song in the AI ocean: a survey on hallucination in large language models. arXiv preprint arXiv:2309.01219.