Is More Human-Like Better? The Impact of Virtual Streamer Language Naturalness on User Engagement

Abstract

The proliferation of livestreaming eCommerce platforms has witnessed a surge in the utilization of virtual streamers. Language naturalness of streamer refers to how "human-like" the streamer sounds in both their speech (voice naturalness) and scripted language (text naturalness), serving as a crucial informational cue that can swiftly capture user attention. However, empirical investigation into the significance of language naturalness exhibited by virtual streamers within the context of livestreaming events remains largely unexplored. To address this critical yet under-researched issue, this study employs a comprehensive approach, integrating deep learning networks, large language models, fundamental frequency extraction, and econometric analysis on a dataset comprising 394 livestreaming events from a large livestreaming platform. Building on the information foraging theory, this study examines the impact of both text and voice naturalness on user engagement. The findings reveal that text naturalness exerts a negative influence on user engagement, whereas voice naturalness yields a positive impact. Moreover, anthropomorphic appearance (i.e., appearance resembling human-like features) mitigates the negative effect of text naturalness and amplifies the positive effect of voice naturalness on user engagement. Additionally, the impact of language naturalness on user engagement exhibits heterogeneity.

Keywords: Livestreaming eCommerce; Virtual Streamer; Language Naturalness; Cross-Modal Effects; Anthropomorphic Appearance

1. Introduction

Livestreaming, a multimedia entertainment interaction that relies on the internet, has gone viral globally since 2011. As it fused with social media and became an influential marketing resource, it led to the concept of livestreaming eCommerce. Livestreaming eCommerce is an audio-visual broadcast where the streamers vividly present products to capture and retain audience attention and promote purchases. Livestreaming platforms also offer real-time chat interfaces, facilitating seamless interaction between the audience and the streamer (Fei et al., 2021; McLean & Osei-Frimpong, 2019). Distinguished from hedonistic livestreaming, livestreaming eCommerce offers utilitarian services, enables users to experience convenience, practicality, value, and cost-effectiveness while watching the livestreaming. In 2023, China's livestreaming eCommerce market reached 4.9 trillion yuan (approximately US$681.7 billion), reflecting a year-on-year growth rate of 35.2% (iResearch, 2024). The global livestreaming eCommerce market is projected to reach US$6 trillion by 2035 (Transparency Market Research, 2025). The above figures highlight the growing appeal of livestreaming eCommerce in attracting audiences worldwide and promoting sales.

Driven by the new factor of production as digital technology (Yang et al., 2023), leveraging the synergistic potential of artificial intelligence (AI), big data, and virtual reality, the landscape of livestreaming eCommerce platforms has witnessed a surge in virtual streamers—AI-driven digital personas or animated characters that are autonomously generated and operated by algorithms. In early 2020, Taobao—a leading online eCommerce platform in China operated by Alibaba Group and known for its vast range of consumer goods and livestreaming commerce features—proposed prioritizing supporting virtual

streamers in live sales at the “Taobao Live MCN (Multi-Channel Network; the company that integrate and manage content creators such as streamers) Agency Conference.” In recent years, several renowned brands such as Dove, L’Oreal, Pepsi, and Winona have ventured into the digital realm by introducing virtual streamers and initiating their operations. This trend continues to accelerate, with the annual growth rate of virtual streamers in livestreaming eCommerce projected to reach 127.89% between 2021 and 2025 (Leadleo, 2021). By 2025, market penetration is expected to reach 39%, with an estimated market size of 26.971 billion yuan (approximately US$3.75 billion; Leadleo, 2021). Furthermore, consumers demonstrate strong receptivity toward virtual streamers as product promoters. A survey conducted by iiMedia Research (2023), based on interviews with 1,654 consumers, reports that nearly 90% of respondents are willing to purchase products promoted by virtual streamers, and 36.7% express an increased likelihood of doing so in the future.

The progression of digital technologies has driven the evolution of virtual streamers beyond traditional virtual agents. As an emerging technology, virtual streamers exhibit distinct advantages in intelligent interaction, particularly through real-time engagement and multimodal coordination. By integrating speech output with anthropomorphic visual representations and synchronizing linguistic and visual elements, virtual streamers enhance user attraction and interactivity in livestreaming environments (Ma et al., 2024). Furthermore, their role has expanded from task-based functionalities, such as information retrieval and task management, to commercial applications, including driving purchases and strengthening brand engagement. This shift necessitates more proactive guidance, enabling virtual streamers to utilize persuasive language and contextual marketing strategies to stimulate user

purchasing intent and create a more immersive shopping experience (Wu et al., 2023). However, these advancements also present notable challenges. In addition to ensuring basic system stability and seamless performance, virtual streamers must meet higher standards for user immersion to achieve highly natural interactions within complex livestreaming environments (Kim et al., 2020). Given these advantages and challenges, it is difficult to directly apply findings from traditional human-computer interaction (HCI) research to the highly interactive and task-oriented context of livestreaming eCommerce. Notably, HCI research in this domain remains largely underexplored, resulting in a significant gap in understanding the potential impact of virtual agents, particularly virtual streamers, on user behavior in livestreaming eCommerce.

In livestreaming eCommerce, language communication serves as a crucial mode of interaction, inherently influencing user engagement. Language naturalness, a crucial information cue adept at swiftly capturing users' attention, shapes their engagement by influencing the perceived quality and credibility of information (Filieri et al., 2018). Merchants aspire for the language used by virtual streamers to closely resemble natural human speech, aiming to preserve the advantages of virtual streamers while effectively substituting for human counterparts, leading to the concept of language naturalness. Given that the notion of language naturalness is distinctive to AI-driven virtual streamers, it is impractical to directly extrapolate research findings from human streamers to their virtual counterparts. However, current research on virtual streaming remains limited to attributes of appearance, attractiveness, and functionality, such as the streamer's sensory language (Hu & Ma, 2023), cuteness, responsiveness (Gao et al., 2023), and integrity (Wang et al., 2024), on

purchasing intentions, with no studies yet having explored the impact of virtual streamers' linguistic characteristics, particularly language naturalness, on user behavior. Therefore, it is imperative to investigate how these dimensions of language naturalness utilized by virtual streamers affect user engagement. It is also necessary to recognize that language naturalness is multifaceted, encompassing both textual and vocal dimensions (Fujisaki, 1997), each potentially yielding distinct effects. Such exploration will offer valuable insights for optimizing the design and operational strategies of virtual streamers.

In addition to conveying information through voice during interactions, virtual streamers rely on visual presentations to communicate with users. The human sensory system operates interdependently (Calvert et al., 2004), meaning that perception from one sensory modality, such as language, can be influenced by another sensory modality, such as vision (Fassnidge & Freeman, 2018; Odgaard et al., 2004). The customer's shopping experience is inherently multisensory, incorporating various perceptual elements within the consumption scenario (Imschloss & Kuehnl, 2019). As a digital shopping guide in livestreaming eCommerce, a virtual streamer's appearance not only influences users' perception of the streamer's identity but also shapes how they interpret the streamer's linguistic information. Consequently, the effect of language naturalness may not operate in isolation but rather in conjunction with visual cues. According to Yao et al. (2024), the anthropomorphism of a virtual streamer's appearance is an important moderating factor in the field of virtual livestreaming. However, its interplay with language naturalness remains underexplored. Therefore, it is essential to investigate this cross-modal correspondence effect by examining the degree of a virtual streamer's appearance anthropomorphism as a moderating variable in

the relationship between language naturalness and user behavior.

To fill the mentioned knowledge gaps, this study investigates the linguistic characteristics of virtual streamers, particularly examining the effects of text naturalness and voice naturalness on user engagement. This study is grounded in information foraging theory (Pirolli & Card, 1999), which offers a theoretical framework for understanding how users search for, interpret, and evaluate information in dynamic environments such as livestreaming platforms (Liu et al., 2023; Xiao et al., 2023). We gathered and analyzed a dataset comprising 394 livestreaming clips featuring virtual streamers employed by Taobao merchants. We employed deep learning networks (NISQA), large language models (BERT Score), and fundamental frequency extraction (Praat) to assess the naturalness of both text and voice. We also developed an econometric model to examine the impact of text and voice naturalness on user engagement. In addition, we also investigated how the virtual streamers' anthropomorphic appearance moderates the relationship between language naturalness and user engagement. The results show that text naturalness has a negative impact on user engagement, while voice naturalness has a positive impact on user engagement. Moreover, anthropomorphic appearance plays a moderating role in the influence of text and voice naturalness on user engagement.

Moreover, the influence of a virtual streamer's language naturalness and anthropomorphic appearance may not be uniform across all consumption contexts. Prior research indicates that users' perceptions and responses in livestreaming environments are shaped by situational variables, such as product price tier, brand reputation, streamer gender, and product category (Niu et al., 2023; Xie et al., 2024; Yao et al., 2024). These contextual

factors shape users' cognitive involvement and expectations, thereby potentially moderating the persuasive impact of linguistic and visual cues. For instance, higher-priced products may trigger more systematic processing, rendering verbal naturalness more consequential, while experience products may shift attention toward affective and anthropomorphic cues. Despite these plausible boundary conditions, the existing research has yet to systematically examine how these contextual elements condition the effects of language naturalness in livestreaming commerce. To address this gap, we explore the heterogeneity in the effects of language naturalness and anthropomorphic appearance across product price levels, brand reputations, streamer genders, and product types, thereby offering a more granular understanding of how these cues operate under different contextual contingencies.

Our research makes valuable contributions to the existing literature in several aspects. Firstly, this study takes the lead in exploring a crucial aspect of communication—language naturalness—within virtual streamers, examining its influence on user engagement, an area that has not been previously studied. Our findings reveal the intricate mechanisms by which language naturalness influences user engagement, with a significant impact observed, especially in scenarios characterized by high product prices, experiential product types, unfamiliar brands, and female streamers. This study holds importance as it sheds light on the nuanced factors driving user engagement in the context of virtual streamers, offering valuable insights for marketers and platform operators. Secondly, this study extends traditional HCI research by integrating virtual streamer language naturalness into the broader discourse on AI-mediated communication. Existing studies in traditional virtual agents have mostly focused on traditional scenarios such as customer service and medical care, and few studies

have concentrated on the livestreaming scenario. By analyzing the synergy between text and voice naturalness, this study reveals the compound effect of them on user behavior in livestreaming, providing a new direction for virtual streaming media interaction design. Third, this study unveils that anthropomorphic appearance plays a crucial role in moderating the relationship between language naturalness and user engagement, thereby enriching our comprehension of cross-modal effects on user engagement in livestreaming eCommerce. The study provides deeper insights into how sensory cues interact to shape user experiences in virtual environments, offering valuable implications for marketers and platform operators aiming to enhance engagement and customer satisfaction. Finally, previous research on virtual streamers has primarily employed questionnaire surveys and scenario experiments to capture subjective experiences from respondents. However, this study represents a pioneering effort in collecting objective livestreaming data and developing an econometric model, thereby expanding the range of research methods applicable in the context of virtual streamer livestreaming. By incorporating objective data analysis techniques, this study offers a more comprehensive understanding of user behavior and engagement dynamics in virtual streamer environments, contributing to advancements in research methodology within this field.

2. Theoretical background

2.1 Livestreaming eCommerce

Livestreaming eCommerce offers consumers an immersive virtual shopping experience, fostering value co-creation between customers and businesses. Central to this experience are the streamers (Guo et al., 2022), who are tasked with delivering comprehensive product information and an enjoyable viewing experience to drive user

engagement and purchase (Chen &Yang, 2023; Fei et al., 2021). Previous literature mainly focused on human streamers and studied the influence of streamer characteristics on user engagement and purchase intention. Existing studies have reported that the streamer's attractiveness (Guo et al., 2022), abilities (Liao et al., 2023), emotions (Lin et al., 2021), and language (Liu et al., 2023; Ma et al., 2023) all affect consumer behavior in livestreaming.

With the efficient integration and development of technologies such as AI and virtual reality, the field of intelligent media is rapidly developing, helping virtual streamers perform actively in various media scenarios (Ma et al., 2024). In the commercial realm, virtual streamers differ from hedonic robots, which primarily focus on providing emotional value to users. Instead, the design of virtual streamers is predominantly task-oriented, focusing on customer service and addressing consumers' inquiries and concerns (Rapp et al., 2021). Since 2020, merchants have increasingly embraced virtual streamers for their numerous advantages, including round-the-clock operation and cost-effectiveness (Ma et al., 2024), thereby emerging as a prominent trend in livestreaming eCommerce.

Existing studies on virtual streamers mainly discussed the influence of sensory language (Hu & Ma, 2023), likability, responsiveness (Gao et al., 2023), emotion (Zhou et al., 2024), and integrity (Wang et al., 2024) on users' purchase intention and engagement, as shown in Appendix A. This line of research has predominantly concentrated on virtual streamers' appearance, appeal, and functionalities, examining consumers' subjective experiences and feedback during virtual streamer interactions. There has also been research exploring the optimal signaling strategies for companies using virtual streamers in different market environments (Yu & Yang, 2024). However, there remains a gap in objectively

measuring the design attributes of virtual streamers, particularly a nuanced analysis of how their language impacts user engagement behavior.

2.2 Streamer language in livestreaming eCommerce

Streamer language refers to the communication tactics employed by streamers to promote products, with the goal of attracting customers and stimulating purchases in livestreaming eCommerce. Existing studies have explored the impact of streamers' language on user engagement. Liu et al. (2023) discussed the positive effects of assertive and directive acts and the adverse effects of expressive acts on sales performance in B2B livestreaming events. Ma et al. (2023) discovered that streamers' language appeals positively influence purchase intention through self-referencing and self-brand congruity. Furthermore, Yang and Wang (2022) conducted a corpus-based analysis, revealing gender differences in lexical choices and pragmatic use of addressing words.

Previous studies on streamers' language have primarily focused on human individuals, yet significant differences exist between humans and virtual entities. In virtual streamer livestreaming, AI generates scripts for product promotion, which are then synthesized into clear and natural voices using text-to-speech technology, following the language generation mechanisms of virtual streamers. Unlike human streamers, for whom natural language use is inherent, virtual streamers rely on AI-generated or pre-designed voice systems. Language naturalness focuses on the expressive properties of digital communication, assessing whether a streamer's language reflects features consistent with human communication habits, which include textual and vocal dimensions such as textual formality, expression diversity, signal quality, and prosody (Ephratt, 2011). Consequently,

language naturalness emerges as a critical factor influencing users' perceptions of information quality and virtual streamer credibility (Filieri et al., 2018). The pursuit of human-like linguistic thus characterizes the naturalness of virtual streamers' language, establishing it as a distinctive attribute that differentiates them from human streamers. Nevertheless, the impact of virtual streamers' language on user engagement remains underexplored, particularly with regard to the critical dimension of naturalness.

Existing linguistic research suggests that human speech signals comprise linguistic and paralinguistic information (Ephratt, 2011). Following linguistic theory, language naturalness is conceptualized as comprising two components: linguistic (textual) information such as vocabulary, grammar, and expressive content, and paralinguistic (vocal) elements such as sound quality, rhythm, and pauses (Ephratt, 2011; Fujisaki, 1997). This provides a clear basis to define language naturalness from both text and voice aspects. Hence, this study explores the impact of virtual streamer language on user engagement by comprehensively analyzing language naturalness from both text and voice perspectives.

2.3 Information foraging theory

Information foraging theory, which draws analogies between information-seeking behavior and animal foraging, seeks to explain how individuals search for and evaluate information. Pirolli and Card (1999) introduced this framework, suggesting that individuals, like animals foraging for food, attempt to maximize their information gain while minimizing search effort. A key component of this theory is the concept of information scent, which refers to the cues that individuals rely on to assess the potential value of information before engaging with it. Strong information scents help users make efficient decisions about which

sources to explore, thereby optimizing their search strategies. Over time, the theory has evolved, with researchers refining its applications and expanding the concept of information scent beyond traditional search tasks to diverse digital environments. For example, information scent has been examined in product search websites (Li et al., 2017), social Q&A communities (Shi et al., 2020), and eCommerce platforms (Liu et al., 2023). Under information foraging theory, users assess "information scent"—proximal cues that indicate the value of distal information sources (Pirolli & Card, 1999). In this context, language signals—such as text, audio, and semantic cues—act as dominant sources of information scent that guide attention and engagement (McCart et al., 2013; Moody & Galletta, 2021; Xiao et al., 2023).

Given the expanding applications of information foraging theory, Table 1 summarizes key studies that have extended its scope across various digital contexts, highlighting the role of information scent in guiding user decision-making.

Table 1 Literature review of information foraging theory

Author(s) (Year)	Context	Information Scent
Li et al. (2017)	Product search websites	Title, abstract, keywords and pictures
Liu et al. (2024)	User-generated videos	Video Prototypicality
Shi et al. (2020)	Social Q&A communities	Informational environment and knowledge products of live courses
Xiao et al. (2014)	Health related online sites	Perceived trustworthiness and usefulness of the media
Liu et al. (2023)	Livestreaming eCommerce	Streamers' attractiveness and product fit
Wang & Wu (2019)	Livestreaming eCommerce	Product interactivity, communication immediacy, and peer cues
Xiao et al. (2023)	Livestreaming eCommerce	Product interpretation duration, popularity cue, herding information
The current study	Livestreaming eCommerce	Language naturalness

In the domain of livestreaming eCommerce, information foraging theory provides a theoretically grounded framework for understanding how consumers process real-time information delivered by virtual streamers. While visual elements may shape initial impressions, they rarely convey core transactional content such as product attributes, pricing, or usage scenarios. Unlike traditional eCommerce settings that rely on static browsing, livestreaming requires viewers to make rapid judgments about the relevance and credibility of continuously evolving verbal content. Therefore, in livestreaming commerce, verbal and vocal expressions serve as the primary carriers of product-relevant information, while appearance functions as a moderating rather than a central cue. Prior studies underscore this distinction. For example, the streamer's explanation of product information, product interactivity, communication immediacy, perceived product fit and perceived product suitability will affect the user's engagement and purchase intention (Liu et al., 2023; Wang & Wu, 2019; Xiao et al., 2023). These studies point out that some features of the language aspects of the streamer act as important sources of information scents.

Language naturalness, as an expressive property of speech, may play a pivotal role in processing information in livestreaming. This study extends the application of information foraging theory by examining how virtual streamers' language naturalness functions as an information scent that influences user engagement during the information-seeking process. By integrating linguistic cues into the framework of information foraging, this research offers new insights into how users navigate and process information in livestreaming eCommerce, ultimately advancing our understanding of digital consumer behavior.

2.4 Virtual Voice Agent

In the field of HCI, virtual voice agents are AI-powered programs that utilize natural language processing (NLP) to simulate human-like conversations through text or voice (Shum et al., 2018). The effectiveness of virtual agents is largely shaped by the quality of their verbal and vocal outputs. From a textual perspective, elements such as emotional content, and informality influence users' affective and cognitive evaluations (Ta et al., 2020; Shi et al., 2020). Emotionally expressive language can help curtail loneliness, thus foster user satisfaction and trust (Ta et al., 2020), while excessively informal speech may undermine perceptions of professionalism (Shi et al., 2020). From a vocal perspective, features such as prosody and personalization further shape user perceptions (Araujo, 2018; Kim et al., 2020; Ashktorab et al., 2019). Expressive intonation and smooth rhythm can foster trust and engagement, while robotic cadence or awkward silences can produce discomfort and disconnection(Araujo, 2018; Ashktorab et al., 2019). Personalized vocal qualities that resonate with user expectations can further strengthen satisfaction (Kim et al., 2020). Collectively, these linguistic features act as critical information scents that influence engagement and behavioral outcomes in livestreaming environments.

Among these attributes, language naturalness is particularly influential in shaping user interactions with virtual voice agents. Prior research suggests that language naturalness affects key behavioral and perceptual outcomes in HCI, including social engagement, user acceptance, perceived pleasantness, and compliance. Table 2 summarizes existing studies examining the impact of language naturalness on user interactions with virtual voice agents.

Table 2 Literature review of virtual agents' language naturalness

Author(s) (Year)	Research subject	Contents	Method
Velner et al. (2020)	Robot	Language naturalness → subjective conversational naturalness and social engagement	Experiment
Kühne et al. (2020)	Synthesized voices	Language naturalness →acceptance and likability of the voice	Experiment
Schreibelmayr and Mara (2022)	Speech interfaces	Language naturalness →acceptance, pleasantness and eeriness	Experiment
Becker et al. (2025)	Robot	Language naturalness →compliance and reliance	Experiment
The current study	Virtual streamer in livestreaming	Language naturalness →user engagement	Econometric model

With the continuous advancement of HCI technology, virtual voice agents have been widely adopted across various domains, including customer service, healthcare, education, and social interaction (Brandtzaeg & Følstad, 2017). In recent years, livestreaming eCommerce has emerged as a rapidly growing consumer model, with AI-driven virtual streamers increasingly assuming roles in product promotion and real-time interaction. These virtual streamers exhibit distinctive characteristics, such as high interactivity and task orientation, which set them apart from traditional virtual agents (Ma et al., 2024). Despite the growing presence of virtual streamers in eCommerce, research on HCI within livestreaming contexts remains limited. In particular, the role of virtual agents' language naturalness in livestreaming eCommerce has yet to be explored. Furthermore, while prior HCI studies have predominantly relied on experimental research designs (Becker et al., 2025; Schreibelmayr & Mara, 2022), this study extends HCI research by leveraging real-world consumer data from livestreaming eCommerce environments. Specifically, it examines the impact of both textual and voice naturalness in virtual streamers, offering a deeper understanding of how these linguistic characteristics influence user behavior in commercial livestreaming interactions.

2.5 Cross-modal correspondence effect

The cross-modal correspondence effect posits that human sensory systems do not function in isolation but instead interact with and influence one another. This framework suggests that input from one sensory modality can meaningfully shape perceptions in another (Calvert et al., 2004). Among the various cues that shape user perceptions, voice and visual appearance are especially influential factors (Fink, 2012). Prior research on cross-modal interactions has extensively examined the correspondence between visual and auditory modalities. Much of this research has emphasized the interplay between sound and color features, underscoring their significant effects. For instance, cross-modal associations between sound frequency and color luminance have been shown to draw visual attention (Hagtvedt & Brase, 2016). Similarly, correspondences between sound frequency and color brightness or darkness can shape consumers' willingness to click on products and make purchase decisions (Yang et al., 2022). Furthermore, sensory stimulation employing congruent visual and auditory cues fosters more favorable ethical brand evaluations, which in turn enhances consumers' willingness to pay (Yoganathan et al., 2019).

In the context of cross-modal correspondence, the relative dominance of visual and auditory cues in influencing consumer attention and decision-making remains debated, with growing consensus that such dominance is context-dependent. For instance, Adaval et al. (2019) find that visual information often dominates in traditional advertising settings due to its vivid, emotionally evocative, and associative nature. Visual elements—particularly static or symbolic imagery—tend to capture attention and elicit emotional responses, thereby shaping consumer judgments through intuitive and figurative processing. However, in

interactive digital environments such as livestreaming eCommerce, this pattern does not necessarily hold. Xu et al. (2023), in a large-scale study of livestreaming returns, demonstrate that auditory cues—particularly voice-based expressions—play a more prominent role in driving user behavior, likely due to their immediacy, emotional nuance, and naturalistic delivery in real-time verbal exchanges. These findings underscore that modality dominance is not universal, but contingent upon the interaction setting and task demands.

Unlike conventional voice agents that rely solely on auditory outputs, virtual streamers integrate both auditory and visual modalities through animated digital avatars. As a result, the perceived naturalness of language is not experienced in isolation, but in conjunction with visual anthropomorphic cues, which may reinforce or alter linguistic perception. In livestreaming eCommerce, the primary mode of information transmission—particularly for key product details, pricing, and purchase instructions—is verbal (Liu et al., 2023). Visual features such as facial expressions and gestures typically function in a supplementary role, aiding emotional signaling and user attention but contributing little to semantic content. Especially under the current technological conditions, virtual streamers are highly homogenized in their visual performance, and there is a lack of significant differences between images with similar degree of anthropomorphism. This visual homogeneity constrains the informational richness of appearance-based cues, thereby reinforcing the relative dominance of auditory inputs in shaping user perceptions and judgments.

Nevertheless, prior research suggests that visual input can influence auditory perception in cross-modal processing. Studies have shown that users can align visual and auditory features and that visual stimuli can modulate auditory judgments (Fassnidge &

Freeman, 2018; Odgaard et al., 2004). Such findings provide conceptual grounding for considering visual anthropomorphism as a moderator that may amplify or attenuate the effects of language-based cues on user engagement. Given that the core objective of livestreaming eCommerce is transactional—requiring effective communication of product information—we argue that language remains the dominant modality in shaping user responses. However, visual cues, particularly the anthropomorphic degree of the avatar, may influence how language naturalness is perceived and processed. Consistent with recent findings by Yao et al. (2024), who show that anthropomorphic appearance conditions user evaluations in virtual contexts, we conceptualize visual anthropomorphism as a key moderating variable in our framework. Specifically, we examine how the anthropomorphic appearance of the virtual streamer shapes the cross-modal interplay between linguistic naturalness and user engagement.

3. Research model and hypotheses

3.1 Text naturalness and user engagement

According to information foraging theory, users evaluate the information's value versus the extracting cost based on a particular information scent in order to maximize the gain rate. The user's increased perceived value positively affects the information gain rate, while the increased cost of information acquisition negatively affects the gain rate (Pirolli & Card, 1999). Text naturalness, as a form of information scent, can be categorized into two dimensions—colloquialism and diversity—to evaluate the cost and value of information acquisition, thereby influencing the information gain rate. Colloquial responses may diminish users' perceived value of product information, while diversified expressions can increase the

acquisition cost. Both factors may consequently reduce the user's information gain rate. When users fail to obtain valuable information during livestreaming, they may question the quality and credibility of the product, ultimately decreasing engagement (Filieri et al., 2018).

Specifically, colloquial texts may convey an unprofessional image for merchants (Decock et al., 2021). As the streamer represents the voice of the company, their informal sales script can reduce the persuasiveness of their communication. Consequently, users may perceive the information provided by the streamer as less credible and reliable (Johnson & Grayson, 2005), thereby diminishing the perceived value of the information. Furthermore, livestreaming eCommerce aims to provide a convenient shopping experience and drive higher transaction volumes as a result-oriented utilitarian service (Liu & Xie, 2023). For voice assistants designed for utilitarian services, users typically prefer efficient and beneficial customer interactions (Badghish et al., 2024; Bai et al., 2024; Rapp et al., 2021). However, diversified expressions can increase information complexity, requiring users to expend additional effort to comprehend and process the information (Deng et al., 2021), thereby raising the information acquisition cost. The combined effect of these factors suggests that increased text naturalness is likely to reduce the user's information gain rate during livestreaming, hindering their ability to clearly understand the product and further decreasing user engagement. We thus propose:

H1: *Increased text naturalness in virtual streamers negatively affects user engagement.*

3.2 Voice naturalness and user engagement

The naturalness of a virtual streamer's voice refers to its resemblance to human

speech, representing a critical dimension of language naturalness. In the context of livestreaming, voice naturalness is characterized by appropriate prosody, pauses, and a clear, stable voice signal that closely mimics human speech. Voice naturalness, as a form of information scent, can be categorized into two dimensions: prosody and signal quality. Prosody encompasses elements such as intonation contour, stress patterns, rhythm, and tone of voice (Nicholas Nagel et al., 1994), which enhance the dynamism and emotional expressiveness of the language, making it appear more natural to users (Ehret et al., 2021). Clear speech enables users to perceive the streamer's intonation variations and emphasized points effectively (Scherer, 2003), thereby creating a communication experience that closely resembles human interaction. These dimensions are used to evaluate information acquisition cost and information value, ultimately determining the information gain rate. A voice characterized by rich prosody enhances users' perceived value of product information, rendering the information scent more distinct and engaging. This, in turn, improves information accessibility and credibility. Simultaneously, a clear and stable voice signal reduces information acquisition costs by minimizing users' cognitive load in processing speech content, thereby making information delivery more efficient. The combined effect of these two factors increases users' information gain rate, ultimately enhancing user engagement and interaction (Filieri et al., 2018).

Specifically, appropriate prosodic modulation, well-timed breathing, and pauses enhance the coherence and expressiveness of a dialogue system, resulting in a more natural and rhythmical voice (Adell et al., 2007). Such smooth and natural speech delivery reduces users' discomfort with rigid or mechanical-sounding voices while strengthening emotional

expression, thereby prolonging user interest and increasing their perceived value of information (Cohn et al., 2019). Additionally, high-quality voice signals precisely convey emotions and intonational variations, ensuring clarity, stability, and expressive richness in speech output, which improves information comprehensibility. A clear, well-structured, and expressive voice enables users to better grasp product features, reducing confusion caused by ambiguous or poorly articulated information. This facilitates faster product selection and effectively lowers information acquisition costs (Stevens et al., 2005). Collectively, these factors indicate that voice naturalness enhances users' information gain rate in livestreaming by optimizing the fluency and comprehensibility of information delivery, improving information absorption, and further boosting user engagement and interaction. Thus:

H2: *Increased voice naturalness in virtual streamers positively affects user engagement.*

3.3 Anthropomorphic appearance as a moderator

The degree of anthropomorphic appearance pertains to the extent to which the virtual streamer's appearance closely resembles that of humans. In virtual streamer scenarios, digital humans exhibit more realistic facial features compared to cartoon images, thus possessing a higher degree of anthropomorphic appearance. Visual similarity to humans is another crucial trigger for user anthropomorphic perception, in addition to language (Fink, 2012). Cross-modal correspondence effect suggests that human sensory systems are interdependent, with inputs from one sense significantly influencing perception in other senses in many cases (Calvert et al., 2004). For instance, vision can affect language perception (Sanchez et al., 2006). Therefore, this study considers the anthropomorphic appearance as a moderating

variable, hypothesizing that the anthropomorphic appearance of the virtual streamer will moderate the relationship between the language naturalness and user engagement.

The degree of anthropomorphic appearance will influence users' efficiency in acquiring information while watching livestreaming events. For livestreaming eCommerce, a transaction-oriented utilitarian task, the streamer must be credible because the client needs to believe that the agent will be able to successfully complete the agreed-upon transaction (Gong & Nass, 2007). Humanlike virtual streamers are perceived as more capable and trustworthy. This increase in user-perceived trustworthiness reduces user uneasiness and brings about a reduction in decision-making time (Sanchez et al., 2006), which in turn reduces the cost of information acquisition and thus improves the rate of information acquisition. Considering that virtual streamers with increased text naturalness may decrease user engagement by reducing the rate at which users acquire information, the enhanced information gain rate resulting from a high degree of anthropomorphic appearance could mitigate the negative impact of text naturalness on user engagement. Thus, we propose:

H3: *The impact of text naturalness on user engagement is contingent on anthropomorphic appearance. Specifically, the negative effect of increased text naturalness on user engagement is weaker for virtual streamers with high anthropomorphic appearance than for those with low anthropomorphic appearance*

Virtual streamers with a higher degree of anthropomorphic appearance often exhibit more realistic expressions and detailed gestures to enhance their communication (Ma et al., 2024). These lifelike expressions facilitate the clear conveyance of emotions and intentions, aiding in the expression of thoughts (Krumhuber et al., 2013). Additionally, refined gestures

can complement language by providing additional information and details, while also assisting individuals in constructing and retaining spatial images, thus aiding in the understanding and memorization of complex concepts and information. Consequently, users perceive these virtual agents as more effective communicators (Duffy, 2003). Considering that virtual streamers with a high degree of voice naturalness can enhance user engagement through effective communication skills, an increased anthropomorphic appearance further amplifies the positive impact of voice naturalness on user engagement. This perception prolongs users' interest, leading to a more enjoyable interaction experience between users and virtual streamers. Thus, we propose:

H4: *The impact of voice naturalness on user engagement is contingent on anthropomorphic appearance. Specifically, the positive effect of increased voice naturalness on user engagement is stronger for virtual streamers with high anthropomorphic appearance than for those with low anthropomorphic appearance.*

Figure 1 presents the theoretical model proposed in this study.

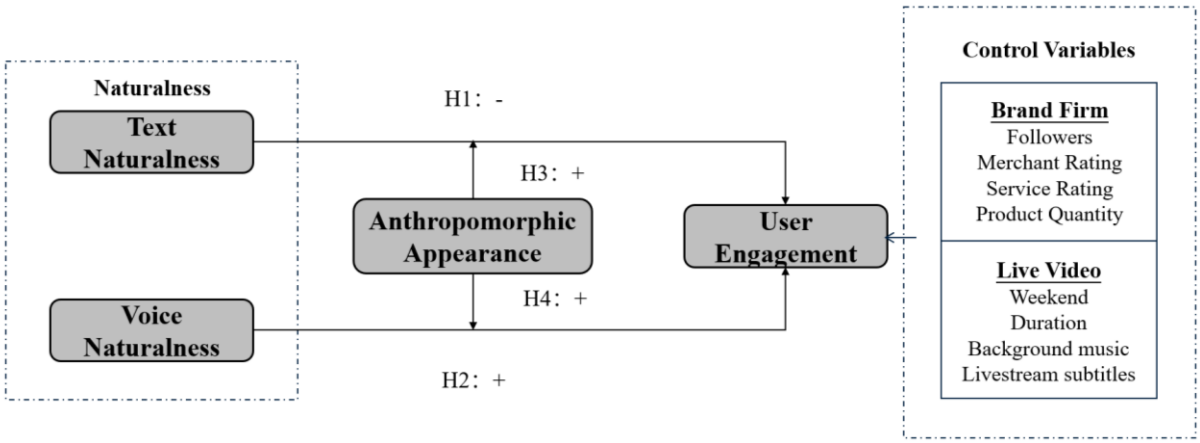


Fig.1 Research Model

4. Research methods

4.1 Data collection

Our empirical setting is Taobao Live, a prominent livestreaming eCommerce platform in China. Established by Alibaba in 2016, Taobao Live has been one of the pioneers in leveraging AI-driven virtual streamer services among eCommerce platforms. It initiated the adoption of virtual streamers in 2020 and stands as one of the leading eCommerce platforms in China. After analyzing the performance of virtual streamers within livestreaming eCommerce, we have summarized their features as follows: 1) They can livestream for extended periods without needing any breaks. 2) Their expressions and movements are more limited compared to humans. 3) They consistently maintain a calm and positive demeanor throughout the livestreaming events. 4) Occasional technical glitches, such as instant teleportation or mismatched lip movements, may occur.

Based on the established criteria, we manually collected data from merchants employing virtual streamers between October 26 and November 1, 2023, across all eight product categories that utilize virtual streamers. In compliance with platform requirements, all merchants using virtual streamers were clearly labeled, enabling the identification of relevant livestreams for data collection. During this period, we randomly selected and recorded the entire livestream from each chosen merchant. This random sampling approach (Xiao et al., 2023) was implemented to minimize potential biases that could arise from selectively choosing livestreams based on content or performance, ensuring that our data sample represented a diverse range of livestreaming sessions. This method also facilitated a balanced distribution across product categories, thereby enhancing the generalizability of our research findings. Additionally, we measured variables for each livestream and aggregated

them to reflect the entire livestreaming period, with each data point corresponding to a unique livestreaming session for a distinct merchant. Additionally, two research assistants independently confirmed whether the streamer was virtual to ensure the accuracy of judgments. We ensured that the retained sample did not include events involving multiple streamers, various product categories, no verbal voice, extremely realistic virtual streamers and those shorter than 1 minute. Consequently, we analyzed 394 livestreaming events, totaling approximately 3709 hours.

4.2 Variables and measures

4.2.1 Independent variables

The main independent variables in this study are text naturalness and voice naturalness. Text naturalness reflects the extent to which language resembles everyday spoken communication, incorporating attributes such as informality, lexical variety, and the use of filler words—together creating an authentic, relaxed tone (Kim et al., 2021). Informal language—marked by colloquial and spontaneous expressions—enhances relatability and reduces perceived scriptedness, thereby increasing perceived naturalness (Dall et al., 2014; Kim et al., 2021). Lexical and topical diversity, reflected in reduced repetition, signals the flexibility and responsiveness characteristic of human dialogue, promoting an engaging and dynamic interaction style (Ram et al., 2018). The inclusion of filler words (e.g., “um,” “you know”) denotes cognitive processing and conversational turn-taking, reinforcing the impression of unscripted and cognitively grounded speech (Bortfeld et al., 2001). Collectively, these features approximate the fluid and adaptive nature of human conversation.

Voice naturalness refers to the clarity and stability of speech, prosodic fluency, and the

use of natural pauses indicative of human cognitive pacing (Ishi et al., 2008). High signal quality enhances intelligibility and reduces perceptual effort, especially in synthetic speech contexts (Stevens et al., 2005). Prosodic variation, including rhythm and intonation, adds expressive nuance and facilitates emotional resonance, while appropriately timed pauses align with human cognitive pacing and turn management (Ehret et al., 2021; Ishi et al., 2008). Together, these auditory features contribute to a conversational speech atmosphere that enhances users' perception of the speaker as fluent, human-like, and trustworthy.

To analyze virtual streamers' sales scripts, we initially converted the livestreaming video to text transcripts using the Lark Converter. Subsequently, two research assistants proofread and revised these transcripts for accuracy. To evaluate the naturalness of virtual streamers' voices, we employed MDX-Net to separate and eliminate background noise, as raw audios often contain elements such as background music, noise, and silence. From the speech transcription, we derived text naturalness metrics such as formality, repetitiveness, and filler words, while voice naturalness metrics like signal quality, prosody, and pauses were obtained from the voice-only audio. Figure 2 provides an overview of the entire measurement process for these two independent variables.

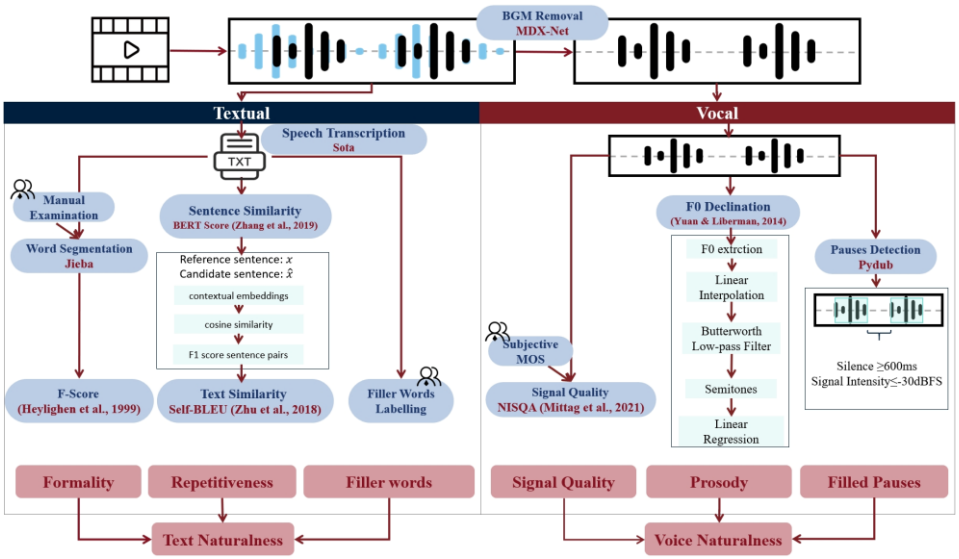


Fig.2 Measurement Process of Independent Variables

(1) Text naturalness

This study employed Francis Heylighen's measure of formality (F-Score), widely utilized for assessing formality in various languages, including English, Dutch, French, Chinese and Italian (Heylighen & Dewaele, 1999). The F-Score indicates that non-deictic words (such as nouns, adjectives, prepositions, and articles) tend to increase with text formality, contrasting with the deictic category (comprising pronouns, verbs, adverbs, and interjections). Formula 1 demonstrates that the F-Score ranges from 0 to 100%, rising with the formality of the text transcript. We utilized the Jieba toolkit for word segmentation and part-of-speech tagging to compute the F-Score for each text. Additionally, to ensure accurate labeling, two research assistants performed manual examinations of all words and corresponding POS tagging.

$$F = (\textit{noun frequency} + \textit{adjective freq.} + \textit{preposition freq.} + \textit{article freq.} - \textit{pronoun freq.} - \textit{verb freq.} - \textit{adverb freq.} - \textit{interjection freq.} + 100)/2 \quad (1)$$

Regarding text repetitiveness, this study extends the applicability of the BERT Score,

originally designed for sentence similarity assessment, to entire paragraphs using the Self-BLEU algorithm. We utilized the Bert-base-Chinese model for contextual embeddings to represent tokens within each sentence, enabling different vector representations for the same word in different sentences based on surrounding context. Employing greedy matching to optimize the cosine similarity score, we calculated the F1 score, a combined measure of precision and recall, for each pair of sentences. Subsequently, we assessed the similarity of each sentence with the rest in a generated set and defined the average score as the level of text repetitiveness. Compared to existing metrics like Type-Token Ratio and BLEU (Papineni et al., 2002), the BERT Score demonstrated a stronger correlation with human evaluations at both the system and segment levels across various common benchmarks in our task.

Filler words, like “I mean,” “you know,” “like,” “uh,” and “um,” are frequently utilized in spoken conversation (Heylighen & Dewaele, 1999). These words serve various functions, including aiding the speaker in thought processing, organizing language, and providing transitions, emphasis, and hints. Consequently, their appropriate usage can enhance the naturalness of synthetic speech (Adell et al., 2007). To analyze their presence, we collaborated with two linguistic scholars to compile and review a list of filler words and calculate their frequency within each transcript.

We transformed formality and repetitiveness into reciprocals due to their negative correlation with text naturalness. Then, we utilized Principal Component Analysis to aggregate formality, repetitiveness, and filler words, providing a comprehensive measure of text naturalness.

(2) Voice naturalness

Previous research has indicated that skilled readers often conclude declarative sentences with noticeable declines in pitch. This decline in pitch at the end of sentences, known as sentence-final pitch (F0) declination, serves as an important linguistic feature in prosody measurement. In this study, we adopted Yuan's method for measuring F0 declination (Yuan & Liberman, 2014). We extracted F0 contours from the audio using Praat with a frame rate of 10 milliseconds for each sentence. To ensure continuity over unvoiced segments, the contours were linearly interpolated and then smoothed using a Butterworth low-pass filter with a normalized cut-off frequency of 0.1. Subsequently, the F0 values were converted to semitones, and linear regression lines were fitted to each semitone using the least-squares method. The slopes of these fitted lines were utilized to represent the F0 declination.

For assessing signal quality, we referred to NISQA, a Deep CNN Self-Attention Model proposed by Mittag et al (2021). Initially, Mel-specs are computed from the input signal, which are then divided into overlapping segments. A framewise neural network is then applied to these Mel-spec segments to compute features suitable for speech quality prediction. In the third stage, the model accounts for the time dependencies of the feature sequence. Finally, the features are aggregated over time in a pooling layer to estimate a single MOS (Mean Opinion Score) value. Given that we used an English dataset (NISQA Speech Quality Corpus), we were concerned that language differences might potentially affect the assessment of voice quality. To validate the accuracy of our deep learning-based signal quality measurement, we supplemented it with Subjective Mean Opinion Scores (MOS). We randomly selected 25% of livestreaming events (100 merchants) and invited two research assistants to manually score the signal quality of streamers' voices. The high consistency

between the two sets of manual scores, indicated by a kappa coefficient of 0.795, demonstrated the reliability of our manual assessment process. Moreover, there was a strong correlation between the mean scores of the two manual sets and the results obtained from NISQA (Pearson's r = 0.842), further confirming the reliability of our computational methodology in scoring signal quality.

The present study utilized the Pydub toolkit to compute the frequency of pauses within each audio segment. A pause was defined as a period of silence exceeding 600ms with a signal intensity less than -30dBFS. To ensure the alignment of these parameter settings with human auditory perception, we manually evaluated the pauses in a small, randomly selected set of audio clips to cross-validate the accuracy of our results.

Finally, we utilized Principal Component Analysis to combine prosody, signal quality and the frequency of pauses, resulting in a measure of text naturalness.

4.2.2 Dependent variable

In our study, the dependent variable is the user engagement in livestreaming. We adopted the standard industry practice of measuring the user engagement level of each livestreaming event by aggregating metrics such as likes, shares, new followers, new members, and bullet screen comments within each event. This comprehensive approach offers a holistic view of user participation during livestreaming (Clement Addo et al., 2021; Zhai & Chen, 2023).

4.2.3 Moderator variable

Yao et al.(2024) demonstrated that virtual streamer anthropomorphism serves as a critical moderating variable in virtual livestreaming. Consistent with their findings, we

classified the anthropomorphism of the streamer's appearance for each livestreaming event into high or low categories. Notably, in our study, each livestreaming event was hosted solely by a single streamer who was represented as a digital human or as a cartoon image. Figures illustrating the streamer appearance are placed in Appendix B, showcasing distinctions in human likeness (high realistic vs. low realistic). We used a binary dummy variable to operationalize this variable, where 1 indicated digital human presenter.

4.2.4 Control variables

To enhance the accuracy of our model estimation, we addressed several potential confounding factors in this study. Specifically, we controlled for confounders at the livestreaming level, including the livestreaming start date (weekdays or weekends), duration, subtitles, and background music, as these factors can influence audience engagement and purchase decisions (Han et al., 2024). Additionally, we accounted for confounders at the merchant level, such as the number of followers, product quantity, and merchant ratings (Liu et al., 2023). Table 3 presents the primary variables and corresponding definitions.

Table 3 Overview of variables in the model

Function	Variable	Operational Definition
Independent variables	Text naturalness	The naturalness of the text conveyed by the virtual streamer
	Voice naturalness	The naturalness of the voice conveyed by the virtual streamer
Dependent variable	User engagement	The number of user engagement during a livestreaming event
Moderating variable	Anthropomorphic appearance	The level of human realism in the streamer's appearance, 1=digital humans, 0=cartoon images
Control variables	Followers	The number of followers of merchants
	Merchant rating	The score of the merchants provided by the Taobao Live platform
	Service rating	The score of the merchants' service by aggregating the degree of service attitude, logistic service and description matching via Principal Component Analysis.
	Duration	The duration of a livestreaming event (minute)
	Product quantity	The number of products sold during a livestreaming event
	Weekend	Dummy variable indicating whether the livestreaming is on a weekend, 1=yes, 0=no
	Subtitles	Dummy variable indicating whether the livestreaming has subtitles, 1=yes, 0=no
	Background music (BGM)	Dummy variable indicating whether the livestreaming has background music, 1=yes, 0=no

5 Results

5.1 Descriptive statistics and correlation

Table 4 provides descriptive statistics for the variables, revealing that the average number of user engagements during the livestreaming was 334.056. The most successful livestreams recorded a total of 8039 user engagements, while some streams garnered no user engagement, demonstrating variability in user engagement across the livestreaming events. Based on the correlation tests conducted (Appendix C), it can be inferred that the variables exhibit relatively low correlations, with a maximum correlation coefficient of 0.422. This suggests that multicollinearity is not a concern in this study.

Table 4 Descriptive statistics

	Mean	SD	Min	Max
User engagement	334.056	926.071	0	8039
Text naturalness	0.378	0.176	0	1
Voice naturalness	0.360	0.262	0.00277	1
Appearance	0.494	0.501	0	1
Followers	13.751	2.126	3.091	17.698
Merchant rating	4.836	0.499	0	5
Service rating	-0.000352	0.360	-0.522	1.323
Duration	6.316	0.808	3.178	9.341
Product quantity (ln)	3.399	1.043	0.693	6.144
Weekend	0.519	0.500	0	1
Subtitles	0.154	0.362	0	1
BGM	0.180	0.384	0	1

5.2 Main models

5.2.1 Model specification

Given that the user engagement variable in our dataset is a non-negative count variable, its variance (926.071) exceeds its mean (334.056), indicating over-dispersion. Additionally, the Pearson dispersion statistic of 1760.983 surpasses the threshold value of 1, further confirming over-dispersion in the dependent variable. These characteristics align with

the assumptions of the negative binomial regression model, as traditional ordinary least squares (OLS) regression would produce biased and inefficient estimates in such cases. To address this issue, we employ a negative binomial regression model with robust standard errors to estimate our research model, following the methodological approach recommended by Gul et al. (2020). The econometric model is specified as formula 2:

$$\begin{aligned} \textit{User engagement} = {} & \beta_0 + \beta_1 \textit{Text naturalness} + \beta_2 \textit{Voice naturalness} + \beta_3 \textit{Text naturalness}^* \\ & \textit{Appearance} + \beta_4 \textit{Voice Naturalness}^* \textit{Appearance} + \beta_5 \textit{Followers} + \\ & \beta_6 \textit{Merchnt rating} + \beta_7 \textit{Service rating} + \beta_8 \textit{Duration} + \beta_9 \textit{Product} \\ & \textit{quantity} + \beta_{10} \textit{Weekend} + \beta_{11} \textit{Subtitles} + \beta_{12} \textit{BGM} + \varepsilon \end{aligned} \quad (2)$$

The dependent variable is livestreaming's user engagement (User engagement). The main independent variables here are text naturalness and voice naturalness, suggesting the characteristics of the language. The moderator variable in this study is anthropomorphic appearance, which interacts with text naturalness and voice naturalness, respectively. Also, we included a set of control variables that could potentially influence the outcome, including the number of followers (Followers), merchant rating (Merchant Rating), service rating (Service Rating), livestreaming duration (Duration), product quantity (Product Quantity), weekend streaming (Weekend), presence of subtitles (Subtitles), and background music (BGM). The coefficients of interest are β_1 (H1), β_2 (H2), β_3 (H3), and β_4 (H4).

5.2.2 Results

The findings from the regression analysis are detailed in Table 5. In Model 1, using control variables only, the impact on user engagement is observed. Model 2, which included independent variables, shows that text naturalness of the virtual streamer has a negative effect on user engagement (β= -1.568, $p < 0.05$) and voice naturalness has a positive effect on user

engagement (β= 0.802, $p < 0.10$). Therefore, hypotheses H1 and H2 are supported. Model 3 include anthropomorphic appearance, for high anthropomorphic streamers, there is a significant positive moderating effect on the relationship between text naturalness and user engagement (β= 2.122, $p < 0.10$), supporting hypothesis H3. The interaction coefficients of voice naturalness and anthropomorphic appearance is also positive and significant (β= 2.191 $p < 0.01$), and thus, hypothesis H4 is supported.

Table 5 The results of main models' estimation

	M1	M2	M3
Followers	0.142*	0.136*	0.148**
Merchant rating	-0.363	-0.182	-0.296
Service rating	0.357	0.066	-0.216
Duration	-0.014	-0.032	-0.015
Product quantity	0.253**	0.147	0.187**
Weekend	0.446*	0.171	-0.362
Subtitles	-0.895***	-0.622*	-0.150
BGM	0.307	0.204	0.010
Text naturalness		-1.568**	-2.483***
Voice naturalness		0.802*	-1.165**
Appearance			0.307
Appearance×Text naturalness			2.122*
Appearance×Voice naturalness			2.191***
Constant	4.516*	4.561**	4.768**

Note: *** represents $p < 0.01$; ** represents $p < 0.05$; * represents $p < 0.10$

5.3 Heterogeneous effects

The existing research on virtual streamers has shown that product price, brands, streamers' gender, and product type influence user engagement behavior (Niu et al., 2023; Xie et al., 2024; Yao et al., 2024). Therefore, we incorporate these factors and explore their heterogeneous effects in our study. We propose that factors such as product price, brands, streamers' gender, and product type may influence users' attention towards the streamers' language. Specifically, price information serves as the primary point of contact for users, playing a critical role in their initial judgment of a product's worthiness for further

exploration. Users tend to prioritize the streamers' elucidation of the product's value proposition in relation to its price (Dodds et al., 1991). Secondly, the establishment of brand reputation can influence users' receptiveness to brand marketing information, including the language used by the streamers. This, in turn, shapes users' expectations and quality judgments of products prior to purchase (Maroufkhani et al., 2022). Thirdly, the gender of streamers may influence the credibility and persuasiveness perceived by users regarding merchants. Hudders and De Jans (2022) propose that gender alignment between streamers and their audiences can enhance persuasion and engagement. Fourthly, in existing studies, products are often categorized as experience products or search products, with the key distinction being whether users can fully comprehend product information (quality and features) before making a purchase. Users must weigh whether they require additional information sources (such as streamers' language) to aid in their purchase decisions (Nelson, 1970). Taking these distinctions into account, we aim to investigate the heterogeneous effects of product price, brand, streamer gender, and product type. We gathered the average product prices and brand reputations from manmanbuy.com, China's leading online shopping product price comparison platform. Additionally, we categorized product types into search products and experience products based on prior research. The findings are presented in Appendix D.

The results show that the correlation between the text naturalness of streamers and user engagement is significant in high price (β= -1.615, $p < 0.05$), niche brand (β= -2.303, $p < 0.05$), female streamers (β= -1.204, $p < 0.10$) and experience products (β= -1.945, $p < 0.05$). In addition, the correlation between the voice naturalness of streamers and user engagement is significant in high price (β= 1.434, $p < 0.05$), niche brand (β= 2.013, $p < 0.01$), female

streamers (β= 1.096, $p < 0.10$) and experience products (β= 1.247, $p < 0.10$). Notably, in the case of low price, famous brand, male voice streamers and search product, the effects of both text naturalness and voice naturalness on user engagement are not significant. In addition, consistent with the main model, anthropomorphic appearance positively moderates the relationship between text naturalness and user engagement and positively moderates the relationship between voice naturalness and user engagement in the case of high price, niche brand, female streamers and experience products. One plausible explanation is that in instances involving low prices, famous brands, male streamers, and search products, user engagement may be less reliant on the streamer's language. Consequently, this diminishes the significance of the relationship between language naturalness and user engagement.

When users face low-risk purchasing decisions, such as buying low-priced products, they generally rely less on the professional knowledge and technical support provided by streamers (Weitz et al., 1986). Low-priced products are typically categorized as low-involvement products, which consumers perceive as less important or personally relevant (Zaichkowsky, 1985). As a result, the decision-making process for these products is often simplified, with consumers prioritizing convenience and speed over a detailed evaluation of the information presented by the streamer. Consequently, the influence of a streamer's language on users' deliberations diminishes when it comes to low-priced products.

In addition, niche brands often suffer from limited market recognition, leaving consumers unfamiliar with aspects such as product quality, performance, and after-sales service. This creates significant information asymmetry, thereby increasing the uncertainty in decision-making. Streamers, acting as information intermediaries, can play a crucial role in

bridging this gap (Nelson, 1970). Niche brands typically emphasize product differentiation and functional value to carve out a competitive edge in the market by focusing on attributes like uniqueness, professional performance, cost-effectiveness, or premium service. Effectively conveying these value propositions requires clear and impactful communication from streamers. When streamers fail to communicate these attributes effectively, consumers' understanding of the product is impaired, which can hinder the development of purchase intentions (Mangold & Faulds, 2009).

Furthermore, the Gender Similarity Theory posits that when salespeople and consumers share the same gender, it fosters emotional resonance and trust, ultimately increasing persuasive effectiveness and boosting purchase intention (Hudders & De Jans, 2022). Since a significant share of users on livestreaming eCommerce platforms are female, these consumers are often more receptive to, and trusting of, female streamers. This is because female consumers perceive female streamers' experiences and product needs as being more aligned with their own, which reduces perceived decision-making risks. However, when male streamers promote products that are primarily targeted toward female audiences—such as clothing, beauty, or household products—consumers may sense a lack of professional expertise or emotional relatability. This perceived disconnect can decrease user engagement and reduce purchase intentions.

Additionally, the distinct characteristics of search products—where users can evaluate performance or value prior to purchase—further limit the impact of a streamer's language on users' purchase decisions (Hoch et al., 1999). In contrast, experience products, whose quality and performance are evaluated during or after use, place greater importance on attributes such

as streamers' trustworthiness, expertise, and communication style (Nelson, 1970). In these cases, a streamer's ability to establish credibility, convey expertise, and communicate effectively has a more substantial influence on shaping users' product perceptions (Erdem & Swait, 2004). Thus, for low-priced items, established brands, search products, or male streamers, users are less influenced by persuasive messaging or sales tactics. As a result, the linguistic strategies employed by streamers in these scenarios are less likely to significantly affect user engagement or purchase intent.

5.4 Robustness checks

We conducted several robustness checks to validate the findings, and the results are presented in Table 6. Firstly, we employed the Poisson regression model instead of the negative binomial regression model to estimate the effects, and the outcomes are depicted in models 1 and 2. Secondly, an alternative method was utilized to assess text naturalness. Specifically, we replaced the measurement of text repetitiveness (BERT-Score) with Bleu, a metric used to evaluate the diversity of generated text based on word embeddings (Papineni et al., 2002). The corresponding results are reported in models 3 and 4. Thirdly, the method for measuring voice naturalness was also altered. We computed the magnitude of F0 declination by subtracting the final fundamental frequency from the peak fundamental frequency (Miller & Schwanenflugel, 2008). The corresponding results are presented in models 5 and 6. Fourthly, we applied winsorization to exclude potential outliers, capturing the top and bottom 0.5% of user engagement. The outcomes are displayed in models 7 and 8. Fifth, we employed a small-sample method, utilizing a 50% subsample, to further assess the robustness of our findings. The results are presented in Models 9 and 10. All results from the

aforementioned models align with our main empirical findings.

Table 6 The results of robustness checks

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
Text naturalness	-1.837***	-3.217**	-1.008**	-1.309***	-1.573**	-2.488***	-1.554**	-2.486***	-1.906***	-4.295***
Voice naturalness	1.126**	-1.007	1.909***	2.589***	0.818*	-1.150*	0.785*	-1.164**	1.409***	-1.594***
Appearance		-0.141		0.707		0.276		0.307		-0.729
Appearance×Text naturalness		2.608*		2.133*		2.141*		2.140*		3.294**
Appearance×Voice naturalness		2.249***		1.214**		2.192***		2.165***		2.886***
Controls	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included
Constant	1.468	2.005	2.709	2.506	4.572**	4.811**	4.561**	4.756**	0.149	2.001

Note: *** represents $p < 0.01$; ** represents $p < 0.05$; * represents $p < 0.10$

5.5 Sensitivity analysis

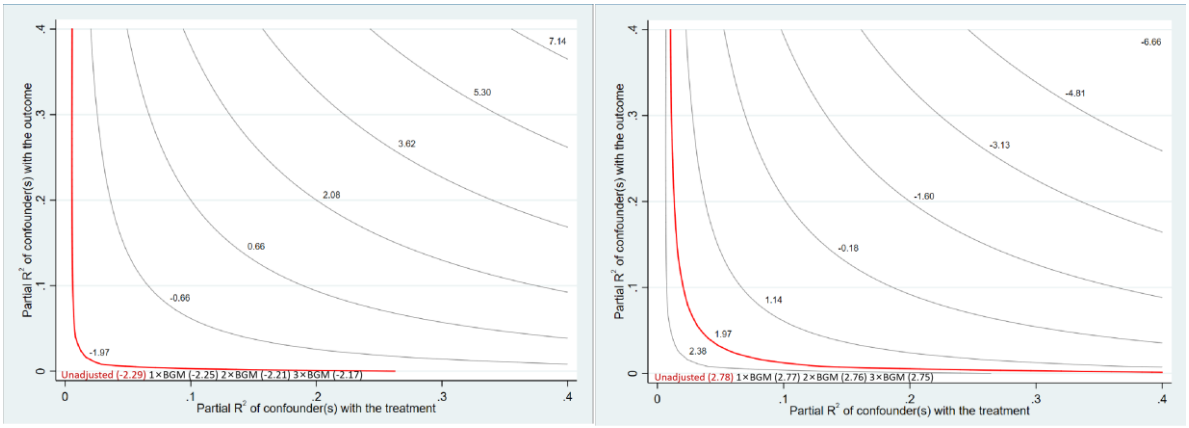
Although our study has undergone a series of robustness tests, the potential threat of omitted variables remains a concern. To address this, we apply the sensitivity analysis method used by Xu et al. (2024) to quantitatively evaluate the impact of omitted variables. The goal of this analysis is to determine the strength of an omitted variable required to undermine the robustness of our current estimation results. In this study, we use the presence of background music (BGM) during livestreaming as the baseline variable. As an auditory input, BGM significantly influences user engagement and is a key determinant in terms of both coefficient magnitude and statistical significance.

The data in the row RV_q of Table 7 indicates that, in the worst-case, omitted variables would need to account for an additional 11.05% and 13.22% of the residual variance simultaneously to completely absorb the influence of the previously identified core explanatory variables. Since both $R^2_{yz.dx}$ and $R^2_{dz.x}$ are lower than RV_q, this suggests that estimates remain robust, even if the potential omitted variables are not as relevant as BGM. Figure 3 illustrates how the results of our sensitivity analysis depend on the strength of the

potential omitted variable. The four data points in the bottom left represent a hypothetical omitted variable with an explanatory power ranging from 0 to 3 times that of the BGM. The t-statistics corresponding to the regression coefficients are also depicted in the figure with the red curve. All four data points fall to the left of the red curve, suggesting that even if the omitted variable has an explanatory power exceeding three times that of BGM, it would not cause the original estimated coefficients to shift from significant to non-significant.

Table 7 Sensitivity analysis of omitted variables.

Baseline Variable	Text naturalness	Voice naturalness
$R^2_{yd.x}$	0.0135	0.0197
RV_q	0.1105	0.1322
$R^2_{yz.dx}$		
1 x BGM	0.0034	0.0034
2 x BGM	0.0069	0.0068
3 x BGM	0.0103	0.0103
$R^2_{dz.x}$		
1 x BGM	0.0014	0.0001
2 x BGM	0.0028	0.0003
3 x BGM	0.0042	0.0004



(a) T-value contour, text naturalness (b) T-value contour, voice naturalness

Fig. 3 Sensitivity analysis contour

5.6 Endogeneity tests

While this study accounts for factors influencing naturalness and user engagement at both the brand-firm and live video levels, certain difficult-to-measure variables, such as

streamer personality and a company's management capacity, may remain unobserved. These factors are associated with text and voice naturalness and may also affect user engagement, potentially leading to endogeneity concerns if not properly controlled. To address this issue, we employed the instrument-free Gaussian copula approach (Park & Gupta, 2012). Prior to implementing this method, we confirmed that the potential endogenous regressors exhibited non-normality (Cramer–von Mises test, ps < 0.05), while the error term followed a normal distribution. The results indicated that the copula terms introduced in the model for the primary explanatory variables (i.e., text naturalness and voice naturalness) are statistically insignificant ($p_{\text{Text naturalness}} = 0.113$, $p_{\text{Voice naturalness}} = 0.995$). Therefore, the study did not uncover any indications of endogeneity issues that could potentially undermine the reliability of our analysis.

Despite conducting multiple robustness and endogeneity tests, potential endogeneity concerns, such as selection bias and reverse causality, may still persist. To further address these concerns, we employed the instrumental variable (IV) approach, which helps mitigate bias arising from omitted variable issues and simultaneity. Specifically, we used the average text naturalness and voice naturalness of similar merchants as an instrument, leveraging variations in these characteristics across comparable merchants to isolate exogenous variation. To identify similar merchants, we computed cosine similarity between each merchant and others based on four key characteristics: followers, brand, merchant rating, and overall experience. These characteristics were selected because they capture critical aspects of merchant positioning and performance within the livestreaming ecosystem. By applying this similarity measure, we matched each merchant with the 100 most comparable peers,

ensuring that our instrument reflects the naturalness levels of merchants operating under similar conditions while remaining exogenous to the focal merchant's specific livestreaming performance. The mean naturalness of similar merchants was selected as an instrumental variable (IV) for two key reasons. First, merchants with similar characteristics are likely to adopt comparable virtual streamer livestreaming strategies, ensuring the relevance of the IV. Since these merchants operate under similar market conditions and competitive pressures, their approaches to text and voice naturalness tend to align, making their average naturalness a strong predictor of the focal merchant's naturalness level. Second, the average naturalness of similar merchants should not directly influence the focal merchant's livestreaming effectiveness (i.e., user engagement), satisfying the exclusion restriction. While merchants within the same peer group may share industry best practices, consumer engagement is primarily driven by the focal merchant's unique livestreaming execution, content quality, and audience interaction. As a result, any variation in similar merchants' naturalness levels affects the focal merchant only through its impact on the focal merchant's own naturalness, rather than through direct spillover effects.

The F-statistics for text naturalness ($F_{Text\ naturalness\_iv} = 73.20$) and voice naturalness ($F_{Voice\ naturalness\_iv} = 47.81$) when treated as potential endogenous variables, both exceed the conventional threshold of 10, rejecting the null hypothesis of weak instruments. The second stage results further demonstrate that the effects of text naturalness ($\beta_{Text\ naturalness} = -11.305$, $p < 0.05$) and voice naturalness ($\beta_{Voice\ naturalness} = 7.465$, $p < 0.05$) on procurement decisions remain significant, consistent with the results of the main regression. In summary, endogeneity is not a serious problem.

6 Discussion and implications

6.1 Discussion of the findings

In recent years, the seamless integration of cutting-edge technologies such as AI, big data analytics, and virtual reality has catalyzed substantial advancements in the intelligent media landscape. Consequently, this has fueled the dynamic performance of AI-driven virtual agents across diverse media environments. Within the realm of livestreaming eCommerce, there is a growing trend among merchants to leverage virtual streamers as a means to showcase products and enhance user engagement and purchase intent. This study underscores the pivotal role played by the language naturalness of virtual streamers in livestreaming eCommerce settings.

Our research findings suggest that the relationship between language naturalness and user engagement is multifaceted and context-dependent, challenging the assumption that more natural language always enhances engagement. Specifically, our results reveal that text naturalness negatively affects user engagement, diverging from previous findings that highlight the benefits of conversational and humanlike language (Kühne et al., 2020; Schreibelmayr & Mara, 2022). This discrepancy stems from the fundamental differences between virtual streamers and the hedonic robots that have been the focus of prior research.

Hedonic social robots are primarily designed to provide companionship, entertainment, and emotional relief. To fulfill these roles, they often employ a humorous, playful, and emotionally expressive linguistic style that fosters an enjoyable and relaxed conversational atmosphere (Liu & Xie, 2023). This approach aligns with user expectations in hedonic contexts, where engagement is driven by entertainment and social interaction rather

than information processing. In contrast, virtual streamers in eCommerce settings serve a predominantly task-oriented function, focusing on the efficient delivery of product-related information. Given that users engage with virtual streamers primarily to obtain product details and make informed purchase decisions, they tend to favor a direct, structured, and efficiency-driven communication style (Rapp et al., 2021). In this context, overly natural or conversational text can introduce inefficiencies in information transmission, leading to a decline in user engagement. Natural language often includes informal expressions, filler words, and tangential content that may distract users and increase cognitive effort, thereby reducing the clarity and immediacy of the intended message. Additionally, a highly conversational style may introduce ambiguity or vagueness, leading to misunderstandings and diminishing trust in the information provided. Users in livestreaming environments expect focused, precise, and professional communication that minimizes the cognitive load associated with processing product information. If the language style deviates from these expectations—by being excessively casual or overly expressive—it can undermine the perceived credibility of both the virtual streamer and the promotional content. As a result, users may disengage due to the perceived mismatch between their expectations and the communication style of the virtual streamer.

Our study advances the theoretical understanding of AI-human interaction by demonstrating the nuanced effects of language naturalness in task-oriented digital environments. While prior research has largely emphasized the benefits of natural and conversational language in hedonic contexts, our findings reveal that these advantages do not necessarily translate to goal-directed interactions such as livestreaming eCommerce. In these

settings, clarity and efficiency take precedence over conversational fluency, and a more structured, less natural linguistic style may be more effective in maintaining user engagement. By identifying the negative impact of text naturalness in task-oriented virtual streaming, we highlight the importance of aligning language styles with the functional demands and user expectations of specific digital interactions. These insights offer critical implications for AI-mediated communication, suggesting that optimizing user engagement requires a context-sensitive approach to language naturalness, rather than a one-size-fits-all application of human-like conversational styles.

Our study reinforces prior findings that voice naturalness positively influences user engagement (Stevens et al., 2005). This relationship suggests that users respond more favorably to virtual streamers whose vocal delivery is authentic, expressive, and engaging, fostering a stronger sense of connection and immersion in the livestreaming experience. Unlike text, which primarily serves an informational function, voice plays a critical role in shaping users' emotional responses and perceptions of social presence. A natural-sounding voice, characterized by appropriate intonation, rhythm, and emotional expressiveness, enhances user trust and attentiveness, ultimately strengthening engagement throughout the interaction. By distinguishing the contrasting effects of text and voice naturalness, our study provides a more nuanced understanding of how different linguistic modalities influence virtual streamer effectiveness. While higher text naturalness can impede information transmission and reduce engagement, greater voice naturalness enhances immersion and fosters a deeper connection between users and virtual streamers. These findings underscore the importance of considering the distinct roles of textual and auditory elements in AI-driven

communication, particularly in task-oriented livestreaming environments.

Furthermore, our findings indicate that the effect of language naturalness on user engagement is contingent upon key situational factors, including anthropomorphic appearance, product price, product type, audience gender, and brand reputation. These factors shape user expectations and preferences, thereby moderating the extent to which language naturalness influences engagement. For instance, highly anthropomorphic virtual streamers may evoke stronger expectations for humanlike communication, amplifying the positive effects of voice naturalness while heightening sensitivity to deviations from expected text styles. Similarly, product characteristics—such as price and type—may determine the optimal balance between clarity and expressiveness. By comprehensively examining these contextual influences, our study advances the understanding of how linguistic and situational factors interact to shape user engagement in virtual streamer interactions. These insights offer practical implications for eCommerce platforms and AI-driven content creators, emphasizing the need for adaptive communication strategies that align with audience expectations and the specific demands of livestreaming scenarios.

6.2 Theoretical implications

This study contributes theoretical insights to the existing literature in several ways. Firstly, this study supplements and extends prior research by investigating the influence of language naturalness exhibited by virtual streamers on user engagement. Language serves as a vital medium for user interaction in livestreaming eCommerce, exerting a significant impact on user engagement. Previous studies have predominantly concentrated on the language characteristics of human streamers (Liu et al., 2023; Ma et al., 2023; Yang & Wang, 2022).

However, given that naturalness is a distinctive attribute of virtual streamers and that users' perceptions of virtual streamers' language may differ from their perceptions of human language, findings regarding the language attributes of human streamers may not be directly transferable to virtual streamers. Our study represents the first empirical investigation into the impact of language attributes displayed by virtual streamers on user engagement behaviors. By shedding light on this previously unexplored area, our findings offer valuable insights into the dynamics of user interactions with virtual streamers in livestreaming eCommerce platforms. This study contributes to the growing body of knowledge in the field by elucidating the nuanced relationship between virtual streamers' language features and user engagement, thus paving the way for future research endeavors in this domain.

Secondly, this study innovatively extends the scope of HCI research by integrating virtual streamers in livestreaming into the broader discourse on AI-mediated communication. Existing research on virtual voice assistants primarily focuses on traditional application scenarios, such as customer service and healthcare (Brandtzaeg & Følstad, 2017), and has not yet examined the impact of HCI on user behavior in livestreaming environments. However, virtual streamers in livestreaming eCommerce exhibit unique interaction characteristics, including real-time and multimodal coordination (Ma et al., 2024; Wu et al., 2023), rendering traditional research findings less applicable to this emerging context. As one of the earliest empirical studies to address this issue, this research reveals the synergistic effects of the dual naturalness of text and speech on user behavior in virtual livestreaming eCommerce scenarios. These findings offer novel insights into HCI research and expand the theoretical boundaries of the field.

Thirdly, this study contributes to the understanding of cross-modal correspondence effects between the auditory and visual modalities in the context of livestreaming scenarios. To the best of our knowledge, we are the first to demonstrate that different levels of appearance anthropomorphism can significantly alter the impact of language naturalness on consumer engagement. Our findings not only confirm that perceptions originating from auditory sensory modality (language naturalness) can be influenced by visual modality (anthropomorphic appearance) (Fassnidge & Freeman, 2018; Odgaard et al., 2004), but also reveal the nuanced and complex nature of this interaction. Specifically, we discovered that the anthropomorphic appearance of the streamer suppresses the impact of text naturalness on user engagement while simultaneously enhancing the influence of voice naturalness on user engagement. This comprehensive approach not only advances our understanding of cross-modal correspondence effects on user engagement in livestreaming eCommerce but also enriches the existing literature in this domain by providing novel insights into the interplay between sensory modalities and their implications for consumer behavior.

Finally, this study contributes to the literature on livestreaming eCommerce by introducing a novel research methodology to assess the language naturalness of virtual streamers. While existing studies in livestreaming predominantly rely on consumer perspective surveys or experiments to explore the psychological mechanisms driving viewers' behavioral intentions, these methods often face several limitations. These include the inherent biases associated with self-reporting in survey-based studies, the constraints imposed by small sample sizes in experimental setups, and the absence of business metrics as dependent variables (Bharadwaj et al., 2022). By integrating deep learning networks, large language

models, fundamental frequency extraction, and econometric analysis, we develop a robust methodology to analyze a dataset comprising 394 livestreaming events from a prominent livestreaming platform. This dataset offers a diverse and extensive representation of livestreaming dynamics, allowing for a thorough investigation into the influence of virtual streamers' language naturalness on user engagement. By bridging the gap between theoretical insights and empirical analysis, our study aims to advance our understanding of livestreaming dynamics and their implications for business success in the digital era.

6.3 Practical implications

First, the naturalness of streamers' text and voice plays a critical role in livestreaming eCommerce, highlighting the importance for managers to prioritize these elements when designing sales scripts and synthesized voices for streamers. Our study emphasizes the importance of formal and concise scripts, as well as clear and rhythmic voices, in enhancing user engagement. For instance, managers of livestreaming platforms and merchants can improve the professionalism of streamers' language by integrating formal terminology into the text while simultaneously simplifying the language to reduce the cognitive load associated with user information retrieval. Additionally, the use of a clear and rhythmic voice allows users to perceive richer and more lifelike auditory details, enhancing their anthropomorphic perception and overall engagement. Therefore, livestreaming platforms and businesses should align virtual streamers' speech design and voice performance with product characteristics and target consumer preferences, fostering advancements in text generation and speech synthesis technologies. Such efforts would improve the efficiency of information delivery, increase user engagement, enhance interaction experiences, and contribute to the

growth of livestreaming eCommerce.

Secondly, our findings indicate that anthropomorphic appearance serves as a moderator, mitigating the negative effect of text naturalness on user engagement while enhancing the positive effect of voice naturalness on user engagement. Consequently, companies should avoid adopting a one-size-fits-all approach to language script configuration and voice synthesis, recognizing the importance of tailoring these elements to align with the specific characteristics of the streamer's image. Our research suggests that firms should develop distinct strategies for streamer language naturalness design based on variations in appearance anthropomorphism observed during livestreaming activities. Thus, when crafting the appearance and voice of virtual streamers, an integrated audio-visual perspective should be adopted to ensure seamless compatibility between facial features and vocal delivery. Livestreaming platforms can offer targeted tools and templates based on the varying anthropomorphic appearances of virtual streamers, assisting businesses in achieving a deep alignment between voice, script, and streamer image. Furthermore, businesses should select appropriate virtual streamer types according to their brand positioning and product characteristics, and customize language styles that align with audience expectations based on the different appearances of streamers. By adopting this holistic approach, companies can optimize the effectiveness of their virtual streamers, enhance user engagement, and foster more meaningful interactions in the livestreaming ecosystem.

Furthermore, firms should avoid indiscriminately replicating livestreaming scripts from successful competitors. Instead, livestreaming platforms and enterprises should adopt a customized language strategy that aligns with their brand positioning and product

characteristics. A tailored language approach not only reinforces brand identity but also enhances user engagement by fostering a more authentic and immersive interaction. Optimizing language naturalness is particularly critical in livestreaming contexts involving high-priced products, niche brands, female streamers, and experience-oriented products. In these cases, consumers are more attuned to the subtlety and authenticity of a virtual streamer's language, making natural and contextually appropriate communication essential for building trust and engagement. To achieve this, firms should invest in advanced text generation and speech synthesis technologies to enhance language delivery quality. Improving linguistic expression can strengthen perceived credibility, bridge the psychological gap between virtual streamers and viewers, and foster brand affinity.

6.4 Limitations

Our research utilizes data from Taobao Live due to its availability. While Taobao Live is a representative platform in China's livestreaming eCommerce industry, our approach aligns with prior studies (Guo et al., 2024) that rely on data from a single platform or cultural context. To mitigate potential sample bias, we include a diverse set of products spanning eight industry categories. However, we acknowledge that the findings may not be fully generalizable to all livestreaming platforms, particularly those with distinct characteristics or cultural contexts. For instance, in high uncertainty avoidance societies, individuals tend to exhibit greater anxiety toward new technologies and ambiguous situations (Hofstede, 1984). Consumers in these markets may be more sensitive to privacy and security concerns related to virtual streamer technologies and more apprehensive about risks such as misinformation, algorithmic manipulation, and deepfake technology. These concerns could lead to

engagement behaviors that differ from those observed in low uncertainty avoidance societies. Future research can enhance the generalizability of our findings by incorporating data from a wider range of livestreaming platforms. A broader dataset would provide deeper insights into the factors shaping user engagement across different cultural and technological landscapes, contributing to a more comprehensive understanding of the livestreaming eCommerce ecosystem. Secondly, this study concentrates on examining the naturalness of both text and voice in virtual streamers. However, the naturalness of other streamer attributes, including facial expressions, gestures, postures, and other visual factors, remains unexplored. Future research endeavors could delve into these supplementary visual features to gain deeper insights into the impact of virtual features on user engagement. Additionally, it is essential to acknowledge that this study did not elucidate the internal mechanisms involved in the process through experimental methods. In future studies, the utilization of experimental methods could provide a means to comprehensively understand the intermediary mechanisms at play. By incorporating experimental approaches, researchers can gain a more nuanced understanding of how various factors interact to influence user engagement in livestreaming.

Reference

Adaval, R., Saluja, G., & Jiang, Y. (2019). Seeing and thinking in pictures: A review of visual information processing. *Consumer Psychology Review*, *2*(1), 50-69.

Adell, J., Bonafonte, A., & Escudero, D. (2007). Filled pauses in speech synthesis: Towards conversational speech. In *International Conference on Text, Speech and Dialogue,* Plzen, Czech Republic.

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*(1), 183–189.

Ashktorab, Z., Jain, M., Liao, Q. V., & Weisz, J. D. (2019). Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). ACM, New York.

Badghish, S., Shaik, A. S., Sahore, N., Srivastava, S., & Masood, A. (2024). Can transactional use of AI-controlled voice assistants for service delivery pickup pace in the near future? A social learning theory (SLT) perspective. *Technological Forecasting and Social Change*, *198*(1), 122972.

Bai, S., Yu, D., Han, C., Yang, M., Gupta, B. B., Arya, V., Panigrahi, P. K., Tang, R., He, H., & Zhao, J. (2024). Warmth trumps competence? Uncovering the influence of multimodal AI anthropomorphic interaction experience on intelligent service evaluation: Insights from the high-evoked automated social presence. *Technological Forecasting and Social Change*, *204*(1), 123395.

Becker, D., Braach, L., Clasmeier, L., Kaufmann, T., Ong, O., Ahrens, K., ... & Wermter, S. (2025). Influence of robots' voice naturalness on trust and compliance. *ACM Transactions on Human-Robot Interaction, 14*(2), 1-25.

Bharadwaj, N., Ballings, M., Naik, P. A., Moore, M., & Arat, M. M. (2022). A new livestream retail analytics framework to assess the sales impact of emotional displays. *Journal of Marketing*, *86*(1), 27–47.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. Language and Speech, 44(2), 123–147.

Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. In *Internet Science: 4th International Conference* (pp. 377-392). Springer, Greece.

Calvert, G., Spence, C., & Stein, B. E. (2004). *The handbook of multisensory processes*. MIT press.

Clement Addo, P., Fang, J., Asare, AO, & Kulbo, NB (2021). Customer engagement and purchase intention in live-streaming digital marketing platforms. *The Service Industries Journal*, *41*(11-12), 767-786.

Chen, N., & Yang, Y. (2023). The role of influencers in live streaming e-commerce:

Influencer trust, attachment, and consumer purchase intention. *Journal of Theoretical and Applied Electronic Commerce Research*, *18*(3), 1601-1618.

Cohn, M., Chen, C.-Y., & Yu, Z. (2019). A large-scale user study of an Alexa prize chatbot: Effect of TTS dynamism on perceived quality of social dialog. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue,* Stockholm, Sweden.

Dall, R., Yamagishi, J., & King, S. (2014). Rating naturalness in speech synthesis: The effect of style and expectation. Speech Prosody, 7, 608–612.

Decock, S., De Clerck, B., Lybaert, C., & Plevoets, K. (2021). Testing the various guises of conversational human voice: The impact of formality and personalization on customer outcomes in online complaint management. *Journal of Internet Commerce*, *20*(1), 1-24.

Deng, Q., Hine, M. J., Ji, S., & Wang, Y. (2021). Understanding consumer engagement with brand posts on social media: The effects of post linguistic styles. *Electronic Commerce Research and Applications*, *48*(1), 101068.

Dodds, W. B., Monroe, K. B., & Grewal, D. (1991). Effects of price, brand, and store information on buyers' product evaluations. *Journal of Marketing Research*, *28*(3), 307-319.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3-4), 177-190.

Ehret, J., Bönsch, A., Aspöck, L., Röhr, C. T., Baumann, S., Grice, M., Fels, J., & Kuhlen, T. W. (2021). Do prosody and embodiment influence the perceived naturalness of conversational agents' speech? *ACM Transactions on Applied Perception (TAP)*, *18*(4), 1-15.

Ephratt, M. (2011). Linguistic, paralinguistic and extralinguistic speech and silence. *Journal of Pragmatics*, *43*(9), 2286-2307.

Erdem, T. l., & Swait, J. (2004). Brand credibility, brand consideration, and choice. *Journal of Consumer Research*, *31*(1), 191-198.

Fassnidge, C. J., & Freeman, E. D. (2018). Sounds from seeing silent motion: Who hears them, and what looks loudest?. *Cortex*, *103*(1), 130-141.

Fei, M., Tan, H., Peng, X., Wang, Q., & Wang, L. (2021). Promoting or attenuating? An eye-

tracking study on the role of social cues in e-commerce livestreaming. *Decision Support Systems*, *142*(1), 113466.

Filieri, R., McLeay, F., Tsui, B., & Lin, Z. (2018). Consumer perceptions of information helpfulness and determinants of purchase intention in online consumer reviews of services. *Information & Management*, *55*(8), 956-970.

Fink, J. (2012). Anthropomorphism and human likeness in the design of robots and human-robot interaction. In *Social Robotics: 4th International Conference*, ICSR 2012, Chengdu, China.

Fujisaki, H. (1997). Prosody, models, and spontaneous speech. In *Computing Prosody: Computational Models for Processing Spontaneous Speech* (pp. 27-42). Springer.

Gao, W., Jiang, N., & Guo, Q. (2023). How do virtual streamers affect purchase intention in the live streaming context? A presence perspective. *Journal of Retailing and Consumer Services*, *73*(1), 103356.

Gong, L., & Nass, C. (2007). When a Talking-Face Computer Agent Is Half-Human and Half-Humanoid: Human Identity and Consistency Preference. *Human Communication Research*, *33*(1), 163-193.

Gul, F. A., Krishnamurti, C., Shams, S., & Chowdhury, H. (2020). Corporate social responsibility, overconfident CEOs and empire building: Agency and stakeholder theoretic perspectives. *Journal of Business Research*, *111*(1), 52-68.

Guo, Y., Zhang, K., & Wang, C. (2022). Way to success: Understanding top streamer's popularity and influence from the perspective of source characteristics. *Journal of Retailing and Consumer Services*, *64*(1), 102786.

Guo, Y., Zhang, Y., Goh, K. Y., & Peng, X. (2024). Can social technologies drive purchases in e-commerce live streaming? An empirical study of broadcasters' cognitive and affective social call-to-actions. *Production and Operations Management*, *0*(0),1-21.

Hagtvedt, H., & Brasel, S. A. (2016). Cross-modal communication: Sound frequency influences consumer responses to color lightness. *Journal of Marketing Research*, *53*(4), 551-562.

Han, L., Fang, J., Zheng, Q., George, B. T., Liao, M., & Hossin, M. A. (2024). Unveiling the

effects of livestream studio environment design on sales performance: A machine learning exploration. *Industrial Marketing Management*, *117*(1), 161-172.

Heylighen, F., & Dewaele, J.-M. (1999). Formality of language: Definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*, *4*(1), 1-38.

Hoch, S. J., Bradlow, E. T., & Wansink, B. (1999). The variety of an assortment. *Marketing Science*, *18*(4), 527-546.

Hofstede, G. (1984). Culture's consequences: International differences in work-related values (Vol. 5). Sage.

Hu, H.-h., & Ma, F. (2023). Human-like bots are not humans: The weakness of sensory language for virtual streamers in livestream commerce. *Journal of Retailing and Consumer Services*, *75*(1), 103541.

Hudders, L., & De Jans, S. (2022). Gender effects in influencer marketing: An experimental study on the efficacy of endorsements by same-vs. other-gender social media influencers on Instagram. *International Journal of Advertising*, *41*(1), 128-149.

iiMedia Research. (2023). *Research Report on China's Virtual Anchor Industry in 2023.* Retrieved from https://report.iimedia.cn/repo13-0/43334.html.

Imschloss, M., & Kuehnl, C. (2019). Feel the music! Exploring the cross-modal correspondence between music and haptic perceptions of softness. *Journal of Retailing*, *95*(4), 158-169.

iResearch. (2024). *China Live Streaming E-commerce Industry Research Report in 2023.* Retrieved from https://report.iresearch.cn/report/202402/4316.shtml.

Ishi, C. T., Ishiguro, H., & Hagita, N. (2008). Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *Speech Communication*, *50*(6), 531-543.

Johnson, D., & Grayson, K. (2005). Cognitive and affective trust in service relationships. *Journal of Business Research*, *58*(4), 500-507.

Kim, S., Eun, J., Oh, C., Suh, B., & Lee, J. (2020). Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot. In *Proceedings of*

the 2020 CHI Conference on Human Factors in Computing Systems. ACM, New York.

Krumhuber, E. G., Kappas, A., & Manstead, A. S. (2013). Effects of dynamic aspects of facial expressions: A review. *Emotion Review*, *5*(1), 41-46.

Kühne, K., Fischer, M. H., & Zhou, Y. (2020). The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. Evidence from a subjective ratings study. *Frontiers in Neurorobotics, 14*(1), 593732.

Leadleo. (2021). *Meta-Universe Series Briefing Report: What is the Application Prospect of Virtual Digital Human in E-commerce Broadcast Area?.* Retrieved from https://www.leadleo.com/report/details/61c920c15774731d8f171ff2

Li, M., Tan, C.-H., Wei, K.-K., & Wang, K. (2017). Sequentiality of Product Review Information Provision. *Mis Quarterly*, *41*(3), 867-A867.

Liao, M., Fang, J., Han, L., Wen, L., Zheng, Q., & Xia, G. (2023). Boosting eCommerce sales with livestreaming in B2B marketplace: A perspective on live streamers' competencies. *Journal of Business Research*, *167*(1), 114167.

Lin, Y., Yao, D., & Chen, X. (2021). Happiness begets money: Emotion and engagement in live streaming. *Journal of Marketing Research*, *58*(3), 417-438.

Liu, C., & Xie, L. (2023). Formal versus casual: How do customers respond to service robots' uniforms? The roles of service type and language style. *International Journal of Hospitality Management*, *114*(1), 103566.

Liu, L., Fang, J., Yang, L., Han, L., Hossin, M. A., & Wen, C. (2023). The power of talk: Exploring the effects of streamers' linguistic styles on sales performance in B2B livestreaming commerce. *Information Processing & Management*, *60*(3), 103259.

Liu, L., Sun, X., Fang, J., & Hossin, M. A. (2024). Exploring prototypicality adherence effects on user engagement in user-generated video platforms. *International Journal of Electronic Commerce*, *28*(3), 381-415.

Liu, Z., Li, J., Wang, X., & Guo, Y. (2023). How search and evaluation cues influence consumers' continuous watching and purchase intentions: An investigation of live-stream shopping from an information foraging perspective. *Journal of Business Research*, *168*(1), 114233.

Ma, E., Liu, J., & Li, K. (2023). Exploring the mechanism of live streaming e-commerce anchors' language appeals on users' purchase intention. *Frontiers in Psychology*, *14*(1), 1109092.

Ma, H., Huang, W., & Dennis, A. (2024). *Who Sells Better? Digital Human Presenter Versus Cartoon AI Presenter in E-commerce Live-Streaming.* In *Proceedings of the 57th Hawaii International Conference on System Sciences,* Hawaii.

Mangold, W. G., & Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix. *Business Horizons*, *52*(4), 357-365.

Maroufkhani, P., Asadi, S., Ghobakhloo, M., Jannesari, M. T., & Ismail, W. K. W. (2022). How do interactive voice assistants build brands' loyalty? *Technological Forecasting and Social Change*, *183*(1), 121870.

Mccart, J. A., Padmanabhan, B., & Berndt, D. J. (2013). Goal attainment on long tail web sites: an information foraging approach. *Decision Support Systems*, *55*(1), 235-246.

McLean, G., & Osei-Frimpong, K. (2019). Chat now… Examining the variables influencing the use of online live chat. *Technological Forecasting and Social Change*, *146*(1), 55-67.

Miller, J., & Schwanenflugel, P. J. (2008). A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children. *Reading Research Quarterly*, *43*(4), 336-354.

Mittag, G., Naderi, B., Chehadi, A., & Möller, S. (2021). NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *ArXiv Preprint ArXiv:2104.09494*.

Moody, G. D., & Galletta, D. F. (2015). Lost in cyberspace: The impact of information scent and time constraints on stress, performance, and attitudes online. *Journal of Management Information Systems*, *32*(1), 192-224.

Nelson, P. (1970). Information and consumer behavior. *Journal of Political Economy*, *78*(2), 311-329.

Nicholas Nagel, H., Shapiro, L. P., & Nawy, R. (1994). Prosody and the processing of filler-gap sentences. *Journal of Psycholinguistic Research*, *23*(1), 473-485.

Niu, B., Yu, X., & Dong, J. (2023). Could AI livestream perform better than KOL in cross-

border operations? *Transportation Research Part E: Logistics and Transportation Review*, *174*(1), 1-24.

Odgaard, E. C., Arieh, Y., & Marks, L. E. (2004). Brighter noise: Sensory enhancement of perceived loudness by concurrent visual stimulation. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(2), 127-132.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics,* Philadelphia, Pennsylvania, USA.

Park, S., & Gupta, S. (2012). Handling endogenous regressors by joint estimation using copulas. *Marketing Science*, *31*(4), 567-586.

Pirolli, P., & Card, S. (1999). Information foraging. Psychological Review, 106(4), 643.

Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., Nunn, J., Hedayatnia, B., Cheng, M., Nagar, A., King, I., & Pettigrue, A. (2018). Conversational AI: The science behind the Alexa Prize. ArXiv preprint, arXiv:1801.03604.

Rapp, A., Curti, L., & Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, *151*(1), 102630.

Sanchez, K., Rosenblum, L. D., & Miller, R. M. (2006). Lipread me now, hear me better later: Crossmodal transfer of talker familiarity effects. *The Journal of the Acoustical Society of America*, *120*(5_Supplement), 3248-3248.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1-2), 227-256.

Schreibelmayr, S., & Mara, M. (2022). Robot voices in daily life: Vocal human-likeness and application context as determinants of user acceptance. *Frontiers in Psychology, 13*(1), 787499.

Shi, W., Wang, X., Oh, Y. J., Zhang, J., Sahay, S., & Yu, Z. (2020). Effects of persuasive dialogues: Testing bot identities and inquiry strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–13). ACM, New York.*

Shi, X., Zheng, X., & Yang, F. (2020). Exploring payment behavior for live courses in social

Q&A communities: An information foraging perspective. *Information Processing & Management*, *57*(4), 102241.

Shum, H. Y., He, X. D., & Li, D. (2018). From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering, 19*(1), 10-26.

Stevens, C., Lees, N., Vonwiller, J., & Burnham, D. (2005). On-line experimental methods to evaluate text-to-speech (TTS) synthesis: Effects of voice gender and signal quality on intelligibility, naturalness and preference. *Computer Speech & Language*, *19*(2), 129-146.

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., Decero, E., & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. *Journal of Medical Internet Research, 22*(3), 1-10.

Transparency Market Research. (2025). *Livestream E-Commerce Market.* Retrieved from https://www.transparencymarketresearch.com/livestream-e-commerce-market.html.

Velner, E., Boersma, P. P., & De Graaf, M. M. (2020). Intonation in robot speech: Does it work the same as with people? In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 569-578).

Wang, X., & Wu, D. (2019, June). Understanding user engagement mechanisms on a live streaming platform. In *International Conference on Human-Computer Interaction* (pp. 266-275). Cham: Springer International Publishing.

Wang, K., Wu, J., Sun, Y., Chen, J., Pu, Y., & Qi, Y. (2024). Trust in human and virtual live streamers: The role of integrity and social presence. *International Journal of Human–Computer Interaction*, *40*(23), 8274–8294.

Weitz, B. A., Sujan, H., & Sujan, M. (1986). Knowledge, motivation, and adaptive behavior: A framework for improving selling effectiveness. *Journal of Marketing*, *50*(4), 174-191.

Wu, R., Liu, J., Chen, S., & Tong, X. (2023). The effect of E-commerce virtual live streamer socialness on consumers' experiential value: An empirical study based on Chinese E-commerce live streaming studios. *Journal of Research in Interactive Marketing*, *17*(5),

714-733.

Xiao, L., Lin, X., Mi, C., & Akter, S. (2023). The effect of dynamic information cues on sales performance in live streaming e-commerce: An IFT and ELM perspective. *Electronic Commerce Research*, *23*(1), 1-30.

Xiao, N., Sharman, R., Rao, H. R., & Upadhyaya, S. J. (2014). Factors influencing online health information search: An empirical analysis of a national cancer-related survey. *Decision Support System.*, *57*(1), 417-427.

Xie, J., Wu, H., Liu, K., Cui, Y., & Zhang, X. (2024). Is virtual streamer useful? Effect of streamer type on consumer brand forgiveness when streamers make inappropriate remarks. *Journal of Retailing and Consumer Services*, *79*(1), 1-13.

Xu, W., Zhang, X., Chen, R., & Yang, Z. (2023). How do you say it matters? A multimodal analytics framework for product return prediction in live streaming e-commerce. *Decision Support Systems*, *172*(1), 113984.

Xu, X., Luo, C., Luo, X. R., & Wang, Z. (2024). Examining how emotions affect online audience retention: Empirical evidence from livestreaming electronic commerce platforms. *Information & Management, 61*(7), 104031.

Yang, N., & Wang, Z. (2022). Addressing as a gender-preferential way for suggestive selling in Chinese e-commerce live streaming discourse: A corpus-based approach. *Journal of Pragmatics*, *197*(1), 43-54.

Yang, S., Chang, X., Chen, S., Lin, S., & Ross Jr, W. T. (2022). Does music really work? The two-stage audiovisual cross-modal correspondence effect on consumers' shopping behavior. *Marketing Letters*, *33*(2), 251-276.

Yang, Y., Chen, N., & Chen, H. (2023). The digital platform, enterprise digital transformation, and enterprise performance of cross-border e-commerce—from the perspective of digital transformation and data elements. *Journal of Theoretical and Applied Electronic Commerce Research*, *18*(2), 777-794.

Yao, R., Qi, G., Wu, Z., Sun, H., & Sheng, D. (2024). Digital human calls you dear: How do customers respond to virtual streamers' social-oriented language in e-commerce livestreaming? A stereotyping perspective. *Journal of Retailing and Consumer Services*,

79(1), 1-21.

Yoganathan, V., Osburg, V. S., & Akhtar, P. (2019). Sensory stimulation for sensible consumption: Multisensory marketing for e-tailing of ethical brands. *Journal of Business Research*, *96*(1), 386-396.

Yu, Y., & Yang, Y. (2024). Signaling effects in AI streamers: Optimal separation strategy under different market conditions. *Journal of Theoretical and Applied Electronic Commerce Research*, *19*(4), 2997-3016.

Yuan, J., & Liberman, M. (2014). F0 declination in English and Mandarin broadcast news speech. *Speech Communication*, *65*(1), 67-74.

Zaichkowsky, J. L. (1985). Measuring the involvement construct. *Journal of Consumer Research, 12*(3), 341-352.

Zhai, M., & Chen, Y. (2023). How do relational bonds affect user engagement in e-commerce livestreaming? The mediating role of trust. *Journal of Retailing and Consumer Services*, *71*(1), 103239.

Zhou, Z., Wen, S., Li, T. T., Zhang, X., & Chi, M. (2024). Virtual streamer and destination visitation: An attractiveness transfer perspective. *Journal of Destination Marketing & Management*, *33*(1), 100922.