

How AI-Generated Language Shapes User Trust in Social Media? A Systematic Review

Hasria Riski S.¹, Janisa Pascawati², Suwarni Suaib³

¹Public Administration Study Program, Faculty of Social and Political Sciences, Universitas Perjuangan Republik Indonesia

²Communications Studies Program, Faculty of Social and Political Sciences, Universitas Perjuangan Republik Indonesia

³Politeknik Negeri Media Kreatif Makassar, Indonesia

*Correspondence: hasriariski13@gmail.com

ABSTRACT

The rapid development of Generative Artificial Intelligence (GenAI) has transformed social media communication by enabling AI systems to produce increasingly human-like language and content. Although research on AI-generated communication and user trust has expanded, findings remain fragmented across platforms, communication contexts, and linguistic characteristics. This study systematically synthesizes empirical evidence on the relationship between AI-generated language and user trust in social media communication. A Systematic Literature Review (SLR) was conducted following the PRISMA 2020 guidelines. Literature was retrieved from the Dimensions database in July 2026 using predefined search terms related to Generative AI, language, trust, and social media. Of 205 records initially identified, 54 full-text studies were assessed for eligibility, and 23 empirical studies were included in the final synthesis. The findings show that research is concentrated in recent years, particularly in marketing and digital communication, health communication, and misinformation and synthetic media. Across the studies, trust is associated with recurring linguistic and communicative features, including human-likeness and conversational style, emotional expression and empathy, personalization, formality and readability, narrative and experiential language, semantic and informational characteristics, transparency and disclosure, and explanatory language. The findings indicate that user trust is not determined by AI authorship alone but emerges through the interaction of relational and epistemic cues within specific communication contexts. This review contributes to human–AI communication research by identifying linguistic and communicative mechanisms through which AI-generated social media content shapes perceptions of trust, credibility, and authenticity.

ARTICLE HISTORY

Published September 23th 2026



KEYWORDS

Generative Artificial Intelligence, AI-generated language, linguistic features, user trust, social media communication.

ARTICLE LICENCE

© 2026 Universitas Hasanuddin
Under the license CC BY-SA 4.0



1. Introduction

The rapid advancement of Generative Artificial Intelligence (Generative AI) has fundamentally transformed the way people produce, distribute, and consume information in digital environments. The emergence of Large Language Models (LLMs) such as ChatGPT, Gemini, Claude, and Copilot has enabled artificial intelligence systems to generate coherent, context-aware, and human-like text. As a result, AI has evolved beyond serving merely as a writing assistant to becoming a communication actor capable of producing messages, responding to conversations, fostering interactions, and influencing public opinion across various social media platforms. This development marks a transition from communication that is entirely human-mediated to human, AI communication, in which AI actively participates in the communication process (Hohenstein et al., 2023; Asif & Zhou, 2024; Yaumi et al., 2023). Furthermore, research has shown that users often struggle to accurately distinguish between human-written and AI-generated text, thereby increasing the potential for manipulation and misperception in digital communication (Jakesch et al., 2023; Junaidi et al., 2020; Pratiwi et al., 2026; Anggawirya et al., 2021).

This transformation has profound implications for the dynamics of digital communication. Social media, once dominated by human-to-human interaction, is now increasingly populated by content that is either generated or assisted by AI. In many situations, users can no longer readily determine whether a post, comment, product review, or persuasive

message originates from a human author or an AI system. This ambiguity presents new challenges for communication research, particularly in understanding how users evaluate message quality and develop trust in the information they encounter. Experimental evidence suggests that AI-generated language can even be perceived as more "human-like" than human-written text by leveraging specific linguistic characteristics, while users continue to rely on relatively inaccurate heuristics when judging the origin of a text (Jakesch et al., 2023; Yaumi et al., 2024; Andini et al., 2026; Hafid et al., 2026).

Trust is a fundamental concept in communication studies because it shapes how individuals evaluate source credibility, message authenticity, information reliability, and ultimately their willingness to accept, share, or act upon information. Within social media environments, trust has become increasingly complex, as message evaluation is influenced not only by informational content but also by linguistic characteristics, source identity, platform norms, and the broader communication context. Consequently, understanding how AI-generated language influences trust formation has become an increasingly important issue from both theoretical and practical perspectives. Previous studies have demonstrated that when users know, or even suspect, that a message has been generated by AI, their evaluations of the message sender tend to become more negative, even when the message content remains unchanged. This finding suggests that perceptions of AI authorship play a significant role in shaping interpersonal trust (Hohenstein et al., 2023).

As the adoption of Generative Artificial Intelligence (Generative AI) continues to accelerate, empirical research examining the relationship between AI-generated language and user trust has expanded rapidly. Existing studies have explored how various linguistic characteristics of AI-generated content shape users' perceptions through features such as emotional language, conversational style, human-likeness, transparency, AI disclosure, authenticity, credibility, framing, personalization, and linguistic clarity. Collectively, these studies suggest that user trust is determined not merely by the fact that content is generated by AI, but also by how its linguistic characteristics are perceived and interpreted by users. For example, more natural language, positive emotional expressions, and appropriate levels of anthropomorphism have been shown to enhance perceptions of social closeness and trust, whereas the effects of AI disclosure vary depending on the communication context and users' expectations (Hohenstein et al., 2023; Greilich et al., 2025).

Despite this growing body of research, several important limitations remain. First, most empirical studies focus on only one or two specific linguistic characteristics, providing little comprehensive understanding of the broader range of language features that contribute to trust formation. Second, existing studies span highly diverse communication contexts including health communication, digital marketing, political communication, education, and misinformation, resulting in fragmented findings that are difficult to integrate into a coherent body of knowledge. Third, methodological approaches vary considerably, encompassing experiments, surveys, discourse analysis, content analysis, corpus linguistics, and computational social science. Although this methodological diversity enriches the literature, it also makes it challenging to identify overarching research patterns and develop a comprehensive understanding of how AI-generated language shapes user trust across different social media contexts.

Several previous review studies have examined Generative AI from perspectives such as technology adoption, AI ethics, education, human-AI interaction, and digital misinformation. However, relatively little attention has been devoted to systematically synthesizing empirical evidence on how the linguistic characteristics of AI-generated social media content influence user trust. This gap is particularly significant because language constitutes the primary interface through which humans interact with AI systems. In other words, users do not interact directly with AI algorithms; rather, they engage with the language those algorithms produce. Understanding which linguistic features foster, or undermine, user trust is therefore essential for advancing communication theory in the era of generative AI.

To address this gap, the present study conducts a Systematic Literature Review (SLR) of empirical research investigating the relationship between AI-generated language and user trust in social media communication. Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, this review synthesizes empirical evidence retrieved from the Dimensions database to identify major research trends, dominant communication contexts, and the linguistic features that have been examined in relation to user trust.

Specifically, this review addresses the following research questions:

RQ1. What are the major research trends and contexts in studies examining AI-generated language and user trust in social media communication?

RQ2. What linguistic features of AI-generated social media content have been examined in relation to user trust?

This study contributes to the fields of communication and linguistics by providing a comprehensive synthesis of the linguistic mechanisms underlying trust formation in AI-generated communication. Unlike previous reviews, which have

primarily emphasized technological developments or the broader adoption of AI, this review positions language as the central unit of analysis for explaining how trust is established, maintained, or diminished in AI-mediated communication. By integrating evidence across diverse communication contexts, social media platforms, and methodological approaches, this review offers a more comprehensive understanding of how linguistic characteristics shape users' trust in AI-generated content. The findings are expected to advance theoretical scholarship on human-AI communication while also providing practical implications for AI developers, social media platforms, digital communication practitioners, and policymakers seeking to design AI-mediated communication that is more transparent, credible, and trustworthy.

2. Methodology

2.1 Research Design

This study employed a Systematic Literature Review (SLR) to systematically identify, evaluate, and synthesize empirical evidence on the relationship between AI-generated language and user trust in social media communication. The SLR approach was selected because it provides a transparent and reproducible method for integrating findings across empirical studies and identifying research trends, communication contexts, and linguistic features associated with trust. The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 framework, covering identification, screening, eligibility assessment, and inclusion of studies (Page et al., 2021).

2.2 Search Strategy

The literature search was conducted in July 2026 using the Dimensions database. The search was limited to publications from 2023 to 2026 and filtered by publication type: Article. Dimensions was selected because it provides broad multidisciplinary coverage of scholarly publications in communication, linguistics, social sciences, and artificial intelligence, making it an appropriate database for identifying relevant empirical studies. The search strategy combined four core concepts: Generative Artificial Intelligence, trust, language, and social media. To ensure that only empirical studies were included, review papers, systematic literature reviews, bibliometric analyses, and meta-analyses were excluded using the NOT operator. The complete search query is presented below:

```
("AI-generated language"  
OR "AI-generated text"  
OR "AI-generated content"  
OR "generative AI"  
OR ChatGPT  
OR "large language model*")  
AND  
(trust  
OR credibility  
OR authenticity)  
AND  
(language  
OR linguistic*  
OR discourse  
OR communication)  
AND  
("social media"  
OR Facebook  
OR Instagram
```

OR TikTok
OR Reddit
OR Twitter
OR X)
NOT
("systematic review"
OR "systematic literature review"
OR bibliometric
OR "meta-analysis")

2.3 Eligibility Criteria

Studies were eligible when they examined generative AI or AI-generated communication, were situated in a social media context, addressed trust or related constructs, and presented empirical evidence. An empirical study was defined as an original investigation that collected, analyzed, or systematically examined observable data using quantitative, qualitative, computational, or mixed-methods approaches. AI-generated communication was defined as communicative content produced, substantially generated, or assisted by generative AI or large language models, including text, posts, comments, responses, explanations, recommendations, and similar content. Social media was defined as digital communication platforms that enable users to create, distribute, interact with, or respond to publicly or semi-publicly accessible content.

Table 1. Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
Studies addressing Generative AI/LLMs as the primary object of communication research	Studies in which AI is not the primary research object
Focus on AI-generated language, text, content, or AI-mediated communication	Studies in which AI is used only as an analytical tool
Conducted in a social media context	Studies not conducted in a social media context
Examining trust, credibility, authenticity, trustworthiness, or reliability	Studies that do not address trust or related constructs
Empirical research	Reviews, bibliometric studies, meta-analyses, editorials, or commentaries
Examining human perceptions, interactions, or responses to AI-generated communication	Studies focusing primarily on algorithm development, benchmarking, fine-tuning, or technical evaluation of AI models

2.4 Study Selection and Quality Appraisal

The retrieved records were screened according to the predefined inclusion and exclusion criteria. The search identified 205 records. After title and abstract screening, 64 records were retained for full-text assessment. Ten full texts could not be retrieved, leaving 54 studies for eligibility assessment. Following full-text assessment, 31 studies were excluded based on the eligibility criteria, resulting in 23 studies included in the final review. The complete selection process is presented in the PRISMA flow diagram (Figure 1).

The methodological quality of the 23 included studies was subsequently assessed using the Mixed Methods Appraisal Tool (MMAT) 2018 (Nha HONG et al., 2018). The appraisal covered qualitative, randomized controlled trial, non-randomized, quantitative descriptive, and mixed-methods studies. Each study was first assessed using the MMAT screening questions and then evaluated according to five design-specific criteria. Responses were recorded as Yes, No,

or *Can't tell*. MMAT appraisal was used to identify methodological strengths and limitations and was not used as an exclusion criterion or converted into an overall numerical score.

2.5 Data Extraction and Synthesis

Data extraction was performed using a structured extraction form developed for this review. The extracted information included publication characteristics, research objectives, communication contexts, social media platforms, research methods, sample or data characteristics, trust-related constructs, linguistic features of AI-generated language, principal findings, and research limitations.

The extracted data were analyzed using thematic synthesis, which enabled the integration of findings from empirical studies employing diverse research designs. Similar findings were grouped into broader conceptual themes to identify recurring patterns across the literature. The synthesis focused on identifying major research trends and communication contexts and examining the linguistic and communicative features of AI-generated social media content investigated in relation to user trust. Methodological quality findings were also considered when interpreting the strengths and limitations of the available evidence.

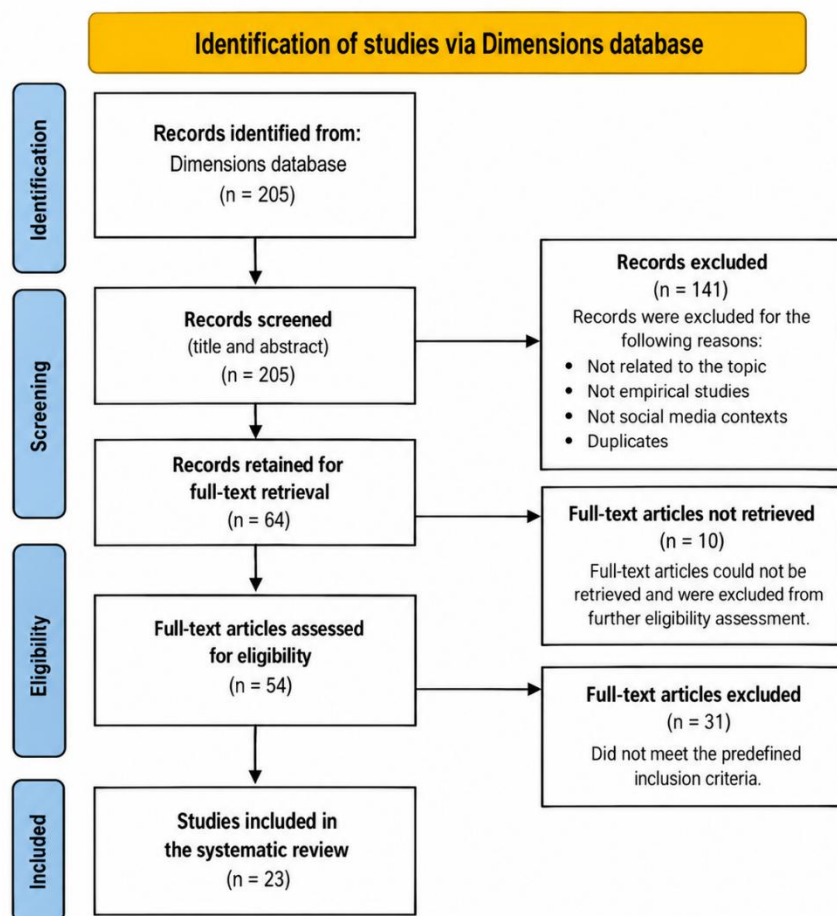


Figure 1. PRISMA 2020 Flow Diagram of the study selection process

3. Results and Discussion

Following the screening and full-text eligibility assessment, a total of 23 empirical studies were included in the final synthesis. To systematically organize and compare the evidence across the included studies, data were extracted using a structured extraction framework. The framework was designed to capture information relevant to both research questions and to accommodate the methodological and contextual diversity of the included studies.

As presented in Table 2, the data extraction framework comprised ten components: (1) article identification, (2) study context and social media platform, (3) research objectives, (4) theoretical or conceptual framework, (5) type of artificial intelligence and form of generated content, (6) linguistic features or language-related aspects, (7) research design

and methods, (8) sample or data, (9) trust-related variables or constructs, and (10) main findings, limitations, and identified research gaps. These components were applied consistently across all 23 included studies.

The extraction framework also provided the basis for answering the two research questions. Information on publication characteristics, study context, social media platforms, research methods, samples, and research settings was used to identify the major research trends and contexts addressed in RQ1. Meanwhile, information on AI-generated content, linguistic features, and trust-related constructs was used to synthesize the linguistic and communicative characteristics examined in relation to user trust in RQ2.

Table 2. Data Extraction Framework for the 23 Included Studies

No.	Extraction Component	Information Extracted
1	Article Identification	Author(s), year, and article title
2	Study Context / Platform	Social media platform and communication context
3	Research Objective	Main purpose or research question of the study
4	Theory or Framework	Theoretical or conceptual framework used
5	Type of AI and Form of Content	AI/LLM used and type of AI-generated communication or content
6	Linguistic Features / Language Aspects	Language-related, linguistic, stylistic, or communicative features examined
7	Research Design and Method	Research design, methodological approach, and analytical method
8	Sample / Data	Participants, posts, comments, datasets, or other empirical data
9	Trust Variables / Constructs	Trust, credibility, authenticity, reliability, trustworthiness, or related constructs
10	Main Findings, Limitations, and Gaps	Principal findings, methodological limitations, and identified research gaps

3.1. Major Research Trends and Contexts in Studies on AI-Generated Language and User Trust in Social Media Communication

The 23 studies included in the final synthesis indicate that research on AI-generated language and user trust in social media communication is a relatively recent and developing field. The included studies were published between 2024 and 2026, with 7 studies published in 2024 (30.4%), 7 in 2025 (30.4%), and 9 in 2026 (39.1%). Thus, 16 of the 23 studies (69.6%) were published in 2025–2026, indicating that empirical attention to AI-generated communication and trust has become increasingly prominent in the most recent years covered by this review. The studies examine diverse forms of AI-mediated and AI-generated communication, including social media posts, conversational responses, product recommendations, explanations, advertising and marketing messages, misinformation rebuttals, and synthetic media.

This recent concentration is consistent with the broader development of research on AI-mediated communication, in which generative AI has increasingly become embedded in online communication environments. Previous research has

highlighted that AI-generated language can be difficult for users to distinguish from human-generated language and that this raises questions concerning credibility, authenticity, and trust in online communication.

Methodologically, the 23 studies demonstrate substantial diversity. The corpus includes mixed-methods studies (8; 34.8%), experimental studies (5; 21.7%), surveys (4; 17.4%), computational or content-analytic studies (3; 13.0%), and qualitative or case-study research (3; 13.0%). This methodological diversity reflects the interdisciplinary nature of the research problem, in which AI-generated communication is examined both as a computational phenomenon and as a form of communication that can influence human perceptions and responses.

Mixed-methods studies combine different forms of evidence to examine AI-generated communication and trust. For example, Saha et al. (2026) combined computational linguistic analysis with qualitative and clinical evaluation of AI- and human-generated mental-health responses, while Lee et al. (2026) combined computational semantic analysis with the examination of AI-generated and human influencer communication. Other mixed-methods studies combined experiments and expert evaluation (Ahmed et al., 2025), questionnaires and interviews (Matsiola et al., 2024), survey and qualitative thematic analysis (Yun & Bickmore, 2025), and social-media data analysis with participant evaluation (Lee et al., 2024). Experimental studies primarily examine how users respond to specific characteristics or conditions of AI-generated communication. These include the effects of AI-generated CSR messages (Illia et al., 2025), LLM-powered bot responses on Twitter (Askari et al., 2024), AI-generated explanations of synthetic media (Ahmed et al., 2025), AI-assisted social media communication (Møller et al., 2026), AI-generated content labels (Li & Yang, 2024), and AI disclosure in social media science communication (Lin & Zhang, 2026).

Computational and content-analytic approaches, in contrast, are particularly evident in studies examining large-scale social media data or the structural and linguistic characteristics of AI-generated content. Bao et al. (2024), for example, examined AI-generated explanations associated with depression-symptom detection, while Lee et al. (2026) analyzed semantic divergence between AI-generated and human influencer recommendations. Malek et al. (2026) used computational methods to analyze misinformation and AI-generated rebuttals, and Sokolova and Yanitskiy (2026) compared human- and AI-generated university promotional posts through content analysis and engagement metrics.

Table 3. Research Designs and Methods of the Included Studies

Research design/method	n	%
Mixed methods	8	34.8
Experimental studies	5	21.7
Surveys	4	17.4
Computational/content analysis	3	13.0
Qualitative/case studies	3	13.0
Total	23	100.0

The platform distribution further demonstrates that research has concentrated on several major social media environments. X/Twitter was represented in 8 studies, followed by Reddit (7), Instagram (6), and Facebook (5). TikTok was represented in two studies, while YouTube, LinkedIn, and VKontakte appeared in one study each. One study examined a simulated Chinese social media environment modeled on Sina Weibo. Because several studies examined multiple platforms, these frequencies are not mutually exclusive.

The prominence of X/Twitter is associated with research on AI-generated news recommendations, misinformation, fact-checking, and AI-generated or synthetic content. Askari et al. (2024) examined LLM-powered bot responses designed to encourage news consumption, while Malek et al. (2026) investigated AI-generated rebuttals in health misinformation. Samet (2025) examined Grok and fact-checking practices on X, and Pfeiffer et al. (2025) analyzed deepfake-driven misinformation distributed through TikTok, Instagram, and X/Twitter. Reddit is particularly prominent in studies examining AI-generated responses and information-oriented communication. Saha et al. (2026) compared AI- and human-written responses in online mental-health communities, Bao et al. (2024) used Reddit datasets for depression-symptom detection

and explanation, Kheiri et al. (2026) analyzed large-scale consumer discourse concerning generative AI, and Lee et al. (2024) examined users' perceptions of ChatGPT-generated health information in Reddit-based kidney-transplant communities. Instagram, meanwhile, is primarily represented in research on marketing, influencer communication, and AI-generated recommendations. Rungruangjit et al. (2024) examined human-like virtual-influencer-generated content, while Lee et al. (2026) compared AI-generated product recommendations with human influencer communication. Instagram also appears in broader studies of social media production and online health-information seeking. Facebook appears primarily in research examining public discourse, AI perceptions, and social media communication. Gallegos (2026), for example, analyzed everyday Facebook discussions about ChatGPT, including posts, comments, memes, and AI-generated images, with particular attention to authenticity, credibility, authorship, and moral evaluation.

Table 4. Social Media Platforms Examined in the Included Studies

Social media platform	Studies (n)
X/Twitter	8
Reddit	7
Instagram	6
Facebook	5
TikTok	2
YouTube	1
LinkedIn	1
Vkontakte	1
Weibo/simulated Chinese social media	1

Note. Platform frequencies are not mutually exclusive because several studies examined more than one platform.

In terms of communication context, the included studies were concentrated in three major domains. Marketing, brand, and digital communication constituted the largest context (8 studies; 34.8%), followed by health communication and information (5 studies; 21.7%) and misinformation, fact-checking, and synthetic media (5 studies; 21.7%). Smaller clusters concerned public discourse and perceptions of AI, education and academic communication, news and media consumption, and general social-media trust and transparency.

Marketing and digital communication studies include virtual influencer communication (Rungruangjit et al., 2024), AI-generated product recommendations (Lee et al., 2026), CSR communication (Illia et al., 2025), generative AI and consumer attitudes (Kheiri et al., 2026), AI-supported social media production (Torres Sáez de Ibarra et al., 2025), university promotion through social media (Sokolova & Yanitskiy, 2026), and AI-supported social media copywriting (Kubovics, 2024). Across these studies, trust-related constructs are commonly connected to authenticity, credibility, reliability, emotional appeal, transparency, and human-likeness.

Health communication represents another substantial research context. Saha et al. (2026) examined AI-generated responses to online mental-health queries; Bao et al. (2024) investigated AI-generated explanations of depression-related social media content; Malek et al. (2026) examined AI-generated rebuttals for health misinformation; Yun and Bickmore (2025) investigated users' use of LLM-based chatbots for health-information seeking; and Lee et al. (2024) examined perceived quality of ChatGPT-generated information concerning kidney transplantation. These studies operationalize trust through constructs such as trustworthiness, credibility, factual accuracy, information quality, authenticity, reliability, and perceived empathy.

A third major cluster concerns misinformation, fact-checking, and synthetic media. Ahmed et al. (2025) examined AI-generated explanations for altered and synthetic images; Malek et al. (2026) examined AI-generated rebuttals to health misinformation; Li and Yang (2024) investigated the effects of AIGC labels on perceived accuracy and message credibility; Lin and Zhang (2026) examined AI disclosure and perceived credibility; Samet (2025) investigated LLM-supported fact-checking; and Pfeiffer et al. (2025) analyzed deepfake-driven misinformation. Collectively, these studies demonstrate that trust-related research extends beyond interpersonal communication to include the credibility and epistemic evaluation of AI-generated information.

Table 5. Major Communication Contexts

Communication context	n	%
Marketing, brand, and digital communication	8	34.8
Health communication/information	5	21.7
Misinformation, fact-checking, and synthetic media	5	21.7
Public discourse and AI perceptions	2	8.7
Education/academic communication	1	4.3
News/media consumption	1	4.3
General social-media trust/transparency	1	4.3
Total	23	100.0

Geographically, the evidence base spans several countries and research settings, including Thailand, China, South Korea, the United States, the United Kingdom, the Philippines, Greece, and India. For example, Rungruangjit et al. (2024) examined consumers in Thailand; Lee et al. (2026) focused on social commerce in South Korea; Illia et al. (2025) examined a UK LinkedIn context; Gallegos (2026) examined Facebook discourse in the Philippines; Matsiola et al. (2024) studied university students in Greece; and Rajkumar et al. (2025) examined university admissions in Southern India. Several other studies used U.S. participants or platform-based datasets, including Askari et al. (2024), Ahmed et al. (2025), Møller et al. (2026), Yun and Bickmore (2025), and Lee et al. (2024).

However, geographical comparison should be interpreted cautiously because not all studies explicitly report a country-level setting, particularly studies based on large platform datasets or multinational online samples. Therefore, the geographical distribution is better interpreted as an indication of where identifiable empirical settings are located, rather than as a complete representation of the global distribution of research.

Overall, the findings reveal three major characteristics of the research landscape. First, research on AI-generated language and user trust is recent and increasingly concentrated in 2025–2026, suggesting a rapidly expanding research area. Second, the field is methodologically heterogeneous, combining computational analysis of large-scale social media data with experiments, surveys, mixed methods, and qualitative approaches. Third, research is concentrated around major social media platforms and high-stakes communication contexts, particularly marketing and digital communication, health information, and misinformation or synthetic media. Across these contexts, trust is not treated as a single construct but is operationalized through related concepts such as trustworthiness, credibility, authenticity, reliability, perceived accuracy, information quality, source credibility, and transparency.

Importantly, the trend analysis also shows that the literature is moving from a broad question of whether users trust AI toward more specific investigations of how characteristics of AI-generated communication shape trust-related judgments. Studies in the corpus examine conversational style, formality, empathy, personalization, emotional language, authenticity, transparency, explanatory quality, readability, human-likeness, semantic structure, and narrative characteristics (Saha et al., 2026; Lee et al., 2026; Møller et al., 2026; Matsiola et al., 2024; Lin & Zhang, 2026; Sokolova & Yanitskiy, 2026).

3.2. Linguistic Features of AI-Generated Social Media Content in Relation to User Trust

The analysis of the 23 studies included in the final synthesis shows that research on AI-generated language and user trust has examined a broad range of linguistic and communicative features rather than a single linguistic characteristic. The features identified across the corpus include human-likeness and conversational style, emotional language and empathy, personalization, formality and readability, narrative and experiential language, semantic and informational characteristics, transparency and disclosure, and explanatory language. These features have been examined in relation to trust-related constructs such as *trustworthiness*, *credibility*, *authenticity*, *reliability*, *perceived information quality*, *source credibility*, and *perceived objectivity*.

3.2.1. Human-likeness and Conversational Style

Human-likeness and conversational style represent recurring features across the reviewed studies. Saha et al. (2026), for example, directly compared the linguistic characteristics of AI- and human-written responses in online mental health communities. AI-generated responses were more verbose, formal, and analytically structured, whereas human responses demonstrated greater linguistic diversity and personal narrative. The study also examined empathy, politeness, personalization, and stance neutrality, linking these characteristics to *trustworthiness*, *authenticity*, *factual accuracy*, and perceived supportiveness.

Similarly, Lee et al. (2026) identified substantial semantic differences between AI-generated product recommendations and human influencer communication. Human influencers relied more heavily on experiential and emotional narratives, whereas AI-generated responses were characterized by systematic conversational structures, functional language, technical terminology, and objective evaluation criteria. These differences were examined in relation to *trustworthiness*, *source credibility*, *authenticity*, and perceived objectivity.

Møller et al. (2026) also examined the perceived quality and authenticity of AI-assisted social media content. The study identified comment length, informativeness, perceived quality, and language perceived as “robotic” or “generic” as relevant characteristics. Participants perceived AI-assisted content as more generic and less personal, which was associated with lower perceptions of authenticity and quality.

Taken together, these findings indicate that human-likeness is broader than simply using informal language. It encompasses conversational naturalness, interpersonal responsiveness, linguistic variation, personalization, and other characteristics that users associate with human communication.

3.2.2. Emotional Language and Empathy

Emotional language and affective expression constitute another recurring linguistic dimension. Rungruangjit et al. (2024) examined emotional and rational appeals in human-like virtual influencer-generated content on Instagram. Emotional appeals included relational and entertainment-oriented content as well as positive and negative emotions, while authenticity and sponsor disclosure were also considered. Relational and entertainment-oriented emotional appeals were more strongly associated with consumers’ willingness to follow and purchase intentions than rational appeals, while negative emotions could also generate engagement.

Saha et al. (2026) found that AI-generated responses demonstrated relatively high cognitive empathy but were less capable of producing deep emotional empathy and personal narratives. This distinction is relevant to trust because users may distinguish between language that merely contains emotional expressions and communication that demonstrates deeper interpersonal understanding.

Emotional and interpersonal language also appears in Gallegos (2026), who examined everyday Facebook discussions about ChatGPT. The study identified humor, irony, satire, confession, politeness, emotional expression, moral evaluation, and vernacular conversational styles as important features through which users negotiated authenticity, legitimacy, credibility, sincerity, and dependence on AI.

Thus, the reviewed studies suggest that emotional language may contribute to trust-related judgments by providing interpersonal and authenticity cues, although its effects are context dependent and interact with other characteristics of AI-generated communication.

3.2.3. Personalization and Interpersonal Address

Personalization is another feature identified across the corpus. Saha et al. (2026) examined personalization through linguistic indicators such as first-person pronoun use and the extent to which responses were adapted to users’ queries. However, AI-generated responses were found to contain less diverse personal narratives than human responses.

Rajkumar et al. (2025) examined personalization in AI-supported digital marketing for student admissions. The study considered personalized advertising alongside emotional sentiment, transparency, and multilingual support, and examined these characteristics in relation to student trust and institutional trust. The findings indicated that personalization was associated with increased student registration, while ethical concerns related to privacy and bias were negatively associated with trust.

Kubovics (2024) similarly examined personalization and sentiment-related characteristics in AI-supported social media copywriting. The study found that users' direct experience and frequency of AI use were associated with higher levels of trust and acceptance toward AI-generated content.

These findings indicate that personalization functions as an interpersonal communication cue. AI-generated content that appears more relevant and tailored to individual users may enhance acceptance and trust, although personalization can also raise concerns about privacy, bias, and transparency.

3.2.4. Formality, Readability, Complexity, and Linguistic Structure

Several studies examined formality, readability, complexity, verbosity, and structural organization as characteristics of AI-generated communication. Saha et al. (2026) explicitly compared verbosity, readability, repeatability, complexity, and formality between AI- and human-written responses. AI-generated responses were more verbose, formal, and analytically structured, while human responses demonstrated greater linguistic diversity and narrative characteristics.

Sokolova and Yanitskiy (2026) similarly compared human- and AI-generated university promotional posts. AI-generated content was characterized by a more neutral, informational, and formal style, whereas human-generated content contained more emotional-evaluative expressions, metaphors, direct personal addresses, and more varied linguistic structures. The study also found differences in engagement and perceived authenticity between the two types of content.

These findings suggest that formal and highly structured language may signal informational competence, but it does not necessarily produce stronger trust. Excessively rigid, generic, or formulaic language may simultaneously reduce perceptions of naturalness and authenticity.

3.2.5. Narrative and Experiential Language

Narrative and experiential language also emerged as an important feature. Lee et al. (2026) found a clear distinction between AI-generated and human influencer communication. Human influencers relied more on experiential language, sensory descriptions, and emotional narratives, whereas AI-generated recommendations relied more heavily on functional language, technical terminology, objective evaluation criteria, and systematic discourse.

Saha et al. (2026) similarly reported that human responses contained greater personal narrative and linguistic diversity than AI-generated responses.

In the context of CSR communication, Illia et al. (2025) examined AI-generated employee testimonials and identified personal anecdotes, balanced positive tone, human-likeness, and non-verbatim repetition as relevant characteristics of AI-generated communication. These characteristics were examined in relation to perceptions of CSR authenticity.

Collectively, these studies suggest that personal narratives and experiential language can function as signals of human presence and authenticity, particularly in communication contexts in which users expect personal experience or testimony.

3.2.6. Semantic, Informational, and Evidence-Based Language

Beyond interpersonal characteristics, several studies focused on semantic and informational characteristics of AI-generated language. Lee et al. (2026) found that AI-generated recommendations were characterized by functional language, technical terminology, objective evaluation criteria, and systematic discourse structures. These characteristics distinguished AI communication from the more experiential and emotionally oriented language used by human influencers.

In health communication, Lee et al. (2024) evaluated ChatGPT-generated responses in terms of information accuracy, authenticity, believability, informativeness, usefulness, and empathy. This demonstrates that users' evaluation of AI-generated language involves not only how a message is expressed but also the informational characteristics conveyed through the language.

Malek et al. (2026) examined AI-generated rebuttals to health misinformation and considered formal versus informal language, emotional narrative, human-versus-AI writing style, coherence, and interpretability at the sentence level. In this context, language functions as a mechanism for communicating corrective information and supporting the credibility of the rebuttal.

The corpus therefore reveals a distinction between interpersonally oriented language and informationally oriented language. The former emphasizes experience, emotion, and relational cues, whereas the latter emphasizes evidence, accuracy, coherence, and informational usefulness.

3.2.7. Transparency, Disclosure, and Authorship Cues

Transparency and disclosure represent another important communication feature. Li and Yang (2024) examined the effects of AI-generated content labels on perceived accuracy, message credibility, and sharing intentions in the context of misinformation. The study directly addressed the role of AI-content labeling as a cue through which users evaluate the credibility and accuracy of information.

Lin and Zhang (2026) examined AI-generated content disclosure in a simulated social media environment and considered conversational style, factual accuracy, warmth, competence, and explanation quality in relation to perceived credibility.

Amina (2026) explicitly investigated the role of AI-generated content transparency in shaping public trust and information acceptance across social media platforms. Transparency was operationalized through disclosure labels, explanatory statements, and metadata indicators, while public trust was assessed through perceived reliability, honesty, credibility, and confidence in AI-generated information.

These findings demonstrate that trust is influenced not only by the linguistic content of an AI-generated message but also by cues that communicate its origin and authorship. Transparency therefore represents a boundary feature between linguistic characteristics and broader communication cues.

3.2.8. Explanatory Language and Interpretability

Explanatory language constitutes another important feature identified in the reviewed literature. Bao et al. (2024) examined natural-language explanations generated by AI models for depression-symptom detection in social media content. The study considered extractive explanations, symptom markers, emotional language, and transparency through explanations of model decisions. Trust, explainability, interpretability, and trustworthiness were assessed through computational metrics and expert evaluation.

Ahmed et al. (2025) similarly examined AI-generated explanations for altered and synthetic images in social media contexts. The explanations included contrastive and teleological forms, with particular attention to narrative structure, clarity, linguistic organization, complexity, and readability. These features were evaluated in relation to trust, credibility, contextual understanding, recall, and opinion stability.

The findings indicate that clear, structured, interpretable, and contextually informative language can function as a trust-related cue, particularly when users need to understand the basis of an AI-generated claim or decision.

3.3. Cross-Study Synthesis, Contradictions, and Research Gaps

The synthesis of the 23 included studies shows that the relationship between AI-generated language and user trust is not determined simply by whether content is produced by AI or humans. Instead, trust-related judgments are shaped by the interaction of linguistic cues, information quality, source cues, and communication context. This finding is consistent with previous research showing that AI-generated language can closely resemble human communication, making the distinction between human- and AI-generated messages increasingly difficult for users (Jakesch et al., 2023).

Across the reviewed studies, human-likeness, conversational style, emotional expression, personalization, authenticity, transparency, and informational quality repeatedly emerged as relevant trust-related cues. However, the findings were not entirely consistent. Some studies associated human-like and emotional communication with authenticity and engagement (Rungruangjit et al., 2024; Gallegos, 2026), whereas others found that AI-generated communication tended to be more formal, structured, and analytical than human communication (Saha et al., 2026; Sokolova & Yanitskiy, 2026). These differences suggest that no single linguistic feature consistently increases trust across contexts. Rather, the relevance of a feature depends on the communication purpose and users' expectations. For example, emotional and

relational language may be more relevant to interpersonal or influencer communication, whereas accuracy, structure, and explanatory quality may be more important in health and misinformation contexts.

A similar inconsistency appears in relation to transparency and AI disclosure. Several studies identify disclosure and labeling as important mechanisms for improving credibility and trust (Li & Yang, 2024; Lin & Zhang, 2026; Amina, 2026). However, disclosure may also increase users' awareness that a message is AI-generated and consequently trigger greater scrutiny. Thus, transparency should not be interpreted as having a uniformly positive effect on trust.

The review also identifies several research gaps. First, most studies examine linguistic features separately, while limited evidence exists on how combinations of features—such as human-likeness, personalization, emotionality, transparency, and factual accuracy—interact to shape trust. Second, trust is operationalized inconsistently across studies through constructs such as *trustworthiness*, *credibility*, *authenticity*, *reliability*, and *information quality*, making direct comparison difficult. Third, most studies focus on single messages, posts, or short-term exposure, leaving limited evidence on how trust develops or changes through repeated and multi-turn interactions. Finally, cross-cultural and multilingual evidence remains limited, particularly regarding whether the same linguistic cues produce similar trust judgments across different cultural and language contexts.

Overall, the findings indicate that trust in AI-generated social media communication is conditional rather than linear. The evidence suggests two broad groups of trust-related cues: relational cues, including human-likeness, conversational style, emotion, empathy, personalization, and narrative; and epistemic cues, including accuracy, informativeness, coherence, objectivity, explanatory quality, and transparency. The main research gap therefore lies in the limited integration of these cues into a unified explanation of how different forms of AI-generated language shape different dimensions of user trust across social media contexts.

4. Conclusion

This systematic literature review synthesized 23 empirical studies examining AI-generated language, social media communication, and user trust. The findings show that research has expanded rapidly in recent years, particularly in marketing and digital communication, health information, and misinformation contexts. X/Twitter, Reddit, Instagram, and Facebook were the most frequently examined platforms.

For RQ2, the review identified eight recurring linguistic and communicative features associated with trust: human-likeness and conversational style, emotional expression and empathy, personalization, formality and readability, narrative and experiential language, semantic and informational characteristics, transparency and disclosure, and explanatory language. These features were associated with trust-related constructs such as *trustworthiness*, *credibility*, *authenticity*, *reliability*, and *information quality*.

The synthesis further indicates that these features can be grouped into two broad categories: relational cues, including human-likeness, emotion, empathy, personalization, and conversational style; and epistemic cues, including accuracy, informativeness, coherence, objectivity, explanatory quality, and transparency. Their relationship with trust is not uniform. Human-like and emotional language may support authenticity and engagement in some contexts, while formal, structured, and evidence-oriented language may be more relevant to credibility and information quality.

The studies also reveal several inconsistencies. For example, human-like communication is sometimes associated with authenticity, but AI-generated language may also be perceived as generic, formal, or less emotionally authentic. Similarly, AI disclosure may improve transparency while simultaneously increasing users' awareness and scrutiny of AI-generated content. These findings suggest that trust is context-dependent rather than a simple positive or negative response to AI-generated communication.

The main research gaps concern the limited examination of interactions among multiple linguistic features, inconsistent operationalization of trust, limited longitudinal and multi-turn studies, and insufficient cross-cultural and multilingual evidence. Future studies should therefore examine how combinations of linguistic cues shape different forms of trust across platforms, cultures, and communication contexts.

Overall, this review shifts the focus from the general question of whether users trust AI toward which linguistic and communicative features make AI-generated communication appear credible, authentic, reliable, or trustworthy, and under what conditions. This perspective positions AI-generated language not merely as a technological output, but as a communicative phenomenon that shapes how users evaluate information and sources in social media environments.

References

- Ahmed, S., Busireddy, T. G., & Dey, S. (2025). Enhancing image comprehension: The impact of AI-generated explanations on perception of altered and synthetic media. In Proceedings of the Eighth AAI/ACM Conference on AI, Ethics, and Society (AIES 2025).
- Amina, N. W. R. (2026). The role of AI-generated content transparency in shaping public trust and information acceptance on social media platforms. *International Journal of Research and Innovation in Applied Science (IJRIAS)*. <https://doi.org/10.51584/IJRIAS.2026.11060270>
- Andini, C., Youngsun, K., Sunnuraini, S., & Warouw, D. S. (2026). Developing a Chunk-Based Learning Model to Improve Korean Speaking Skills for Beginner Learners. *International Journal of L2CT*, 1(2), 156-172.
- Anggawirya, A. M., Prihandoko, L. A., & Rahman, F. (2021, December). Teacher's role on teaching English during pandemic in a blended classroom. In *International Joined Conference on Social Science (ICSS 2021)* (pp. 458-463). Atlantis Press.
- Asif, M., & Gouqing, Z. (2024). Innovative application of artificial intelligence in a multi-dimensional communication research analysis: a critical review. In *Discover Artificial Intelligence* (Vol. 4, Number 1). Springer Nature. <https://doi.org/10.1007/s44163-024-00134-3>
- Bao, E., Perez, A., & Parapar, J. (2024). Explainable depression symptom detection in social media. *Health Information Science and Systems*. <https://doi.org/10.1007/s13755-024-00303-9>
- Gallegos, I. N. (2026). Postcolonial digital reflexivity: Algorithmic coloniality and the Filipino self in everyday Facebook talk about ChatGPT. *AI & Society*. <https://doi.org/10.1007/s00146-026-03122-6>
- Greilich, A., Bremser, K., & Wüst, K. (2025). Consumer Response to Anthropomorphism of Text-Based AI Chatbots: A Systematic Literature Review and Future Research Directions. In *International Journal of Consumer Studies* (Vol. 49, Number 5). John Wiley and Sons Inc. <https://doi.org/10.1111/ijcs.70108>
- Hafid, A., Dalyan, M., Andini, C., Riski, H., Sunnuraini, & Iman, M. N. (2026). When Indonesian is not enough: English as an emotional and pragmatic resource in Indonesian digital communication. *International Journal of Innovative Science and Research Technology*, 11(9), 1737-1742. <https://doi.org/10.38124/ijisrt/26sep655>
- Hohenstein, J., Kizilcec, R. F., DiFranzo, D., Aghajari, Z., Mieczkowski, H., Levy, K., Naaman, M., Hancock, J., & Jung, M. F. (2023). Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, 13(1). <https://doi.org/10.1038/s41598-023-30938-9>
- Illia, L., Ballester-Ripoll, R., & Clausen, A. K. (2025). Fabricating CSR authenticity: The illusory truth effect of CSR communication on social media in the AI era. *Public Relations Review*. <https://doi.org/10.1016/j.pubrev.2025.102588>
- Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences of the United States of America*, 120(11). <https://doi.org/10.1073/pnas.2208839120>
- Junaidi, Hamuddin, B., Julita, K., Rahman, F., & Derin, T. (2020). Artificial intelligence in EFL context: rising students' speaking performance with Lyra virtual assistance. *International Journal of Advanced Science and Technology*.
- Kheiri, P., Raeesi Vanani, I., & Amiri, M. (2026). Designing marketing strategies based on a dual-method analysis of consumer attitudes toward generative AI. *Discover Artificial Intelligence*. <https://doi.org/10.1007/s44163-026-01382-1>

- Kubovics, M. (2024). Experience with artificial intelligence and attitudes towards social media copywriting. University of Ss. Cyril and Methodius in Trnava.
- Lee, J., Park, J., & Han, H. S. (2024). Using ChatGPT for kidney transplantation: Perceived information quality by race and education levels. *Clinical Transplantation*. <https://doi.org/10.1111/ctr.15378>
- Lee, W.-C., Lee, J.-S., & Suh, J. (2026). Semantic divergence in AI-generated and human influencer product recommendations: A computational analysis of dual-agent communication in social commerce. *Applied Sciences*, 16, 2816. <https://doi.org/10.3390/app1602816>
- Li, F., & Yang, Y. (2024). Impact of artificial intelligence-generated content labels on perceived accuracy, message credibility, and sharing intentions for misinformation: Web-based, randomized, controlled experiment. *JMIR Formative Research*. <https://doi.org/10.2196/60024>
- Malek, S., Griffin, C., Fraleigh, R. D., Lennon, R., Monga, V., & Shen, L. (2026). Intervention in health misinformation using large language models for automated detection, thematic analysis, and inoculation: Case study on COVID-19. *Journal of Medical Internet Research*. <https://doi.org/10.2196/75500>
- Matsiola, M., Lappas, G., & Yannacopoulou, A. (2024). Generative AI in education: Assessing usability, ethical implications, and communication effectiveness. *Societies*, 14(12), 267. <https://doi.org/10.3390/soc14120267>
- Møller, A. G., Romero, D. M., Jurgens, D., & Aiello, L. M. (2026). The impact of generative AI on social media: An experimental study. *Scientific Reports*. <https://doi.org/10.1038/s41598-026-40110-8>
- Nha HONG, Q., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., Rousseau, M.-C., & Vedel, I. (2018). MIXED METHODS APPRAISAL TOOL (MMAT) VERSION 2018 User guide. <http://mixedmethodsappraisaltoolpublic.pbworks.com/>
- Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *The BMJ*, 372, n160. <https://doi.org/10.1136/BMJ.N160>
- Pfeiffer, A., Krishna, N., Wernbacher, T., & Seböck, W. (2025). From hoax to reality: Deepfake-driven misinformation and the death of Ozzy Osbourne. University for Continuing Education Krems.
- Pratiwi, N. I., Sosrohadi, S., Rahman, F., Dalyan, M., & Andini, C. (2026). The use of code-mixing by edu-influencers in social media: A communication strategy in sociolinguistic perspective. *International Journal of Research and Innovation in Applied Science*, 11(1), 1034-1042.
- Rajkumar, S. G., Joseph, S., Sudhakar, C., & Jacinth, J. J. (2025). The impact and ethics of artificial intelligence on digital marketing for student admissions in Southern India. *Journal of Information Systems Engineering and Management*, 10(25s).
- Rungruangjit, W., Mongkol, K., Piriyaikul, I., & Charoenpornpanichkul, K. (2024). The power of human-like virtual-influencer-generated content: Impact on consumers' willingness to follow and purchase intentions. *Computers in Human Behavior Reports*, 16, 100523. <https://doi.org/10.1016/j.chbr.2024.100523>
- Saha, K., Jain, Y., Rodriguez, V. J., & De Choudhury, M. (2026). Linguistic comparison of AI- and human-written responses to online mental health queries. *npj Artificial Intelligence*. <https://doi.org/10.1038/s44387-026-00099-x>
- Samet, U. (2025). The positive influence of large language models on fact-checking practices: A case study of Grok. *World Journal of Advanced Engineering Technology and Sciences*, 15(3), 1727–1738. <https://doi.org/10.30574/wjaets.2025.15.3.1123>
- Sokolova, A. A., & Yanitskiy, L. S. (2026). Artificial intelligence for university promotion in social media. *Virtual Communication and Social Networks*, 5(2), 151–161. <https://doi.org/10.21603/vcsn-2026-5-2-151-161>

- Torres Sáez de Ibarra, A., Monteagudo Barandalla, L., & Sánchez Crespo Pérez, I. (2025). Inteligencia artificial en redes sociales: Impacto en la producción audiovisual y las métricas digitales. *VISUAL Review*, 17(3), 21–32. <https://doi.org/10.62161/revvisual.v17.5782>
- Yaumi, M. T. A. H., Rahman, F., & Sahib, H. (2023). Exploring WhatsApp as Teaching and Learning Activities during Covid-19/New Normal era: A Semiotic Technology Analysis. *International Journal of Current Science Research and Review*, 6(12), 7627-7634.
- Yaumi, M. T. A. H., Rahman, F., & Sahib, H. (2024). Bridging Language and Technology through Semiotic Technology. *International Journal of Social Science Research and Review*, 7(1), 52-61.
- Yun, H. S., & Bickmore, T. (2025). Online health information-seeking in the era of large language models: Cross-sectional web-based survey study. *Journal of Medical Internet Research*, 27, e68560. <https://doi.org/10.2196/68560>