IJHSR
ijhighschoolresearch.org
REVIEW ARTICLE

How Large Language Models Are Shaping Online Information Credibility

Ronav A. Lamba

Adlai E. Stevenson High School, 1 Stevenson Dr., Lincolnshire, IL, 60089, USA; ronav.lamba@gmail.com
Mentor: Dr. Siddharth Krishnan, Samuel Lefcourt

ABSTRACT: From homework help to breaking news, AI tools like ChatGPT and Gemini are shaping what many people read and believe online, often without them realizing the source is not a human. The belief and confidence people place in the content's accuracy and reliability are influenced by large language models (LLMs) writing in a human-like style when generating content. This likeness makes the content seem trustworthy, but the overlap of human and machine voices raises concerns about the credibility of digital information. This paper investigates how the spread of LLM-generated content is influencing public trust in online information. It focuses on how people judge credibility when faced with AI-generated material and what factors, such as content labeling, detection tools, platform design, and digital literacy, affect that trust. Instead of running new experiments, this review draws on existing academic research, journalistic work, surveys, and policy reports to examine how trust in AI-generated content is defined, measured, and influenced. The reviewed studies show that content fluency often leads users to overestimate reliability, while education about AI mechanisms and limitations improves their ability to evaluate accuracy. The review further concludes that education, supported by clear accountability and governance policies, is the most consistently supported and scalable approach for restoring calibrated trust in AI-generated content.

KEYWORDS: Computer Science and Software Engineering, Artificial Intelligence, Large Language Models, AI Literacy, Online Information Credibility.

Introduction

Every day, an increasing number of people turn to AI for quick answers, persuasive explanations, and even personal advice. Large language models like ChatGPT and Gemini can produce fluent, convincing language on almost any topic, from education to politics. This review differs from existing work by focusing on LLMs specifically rather than AI systems in general and by examining how their linguistic fluency reshapes credibility judgment. In this review, trust refers to users' willingness to rely on AI-generated content based on perceived accuracy, fairness, and understandability.

This paper examines academic literature, surveys, bibliometric reviews, and policy reports on AI usage, as well as key factors such as content labeling, platform design, detection tools, and digital literacy. Studies show that without AI literacy, users misjudge AI outputs, while stronger literacy improves evaluation and judgment.[1] Henrique and Santos have mapped the rise of academic work on AI transparency and explainability.[2] Freedom House has warned that governments are already using generative AI to spread propaganda.[3] Benk points out that studies on trust often use inconsistent methods and focus on narrow demographics, making it hard to see the whole picture.[4] Adding to the challenge, Huang and Chang argue that LLMs don't actually "reason" but instead mimic patterns of thought, which can make confident-sounding but false statements more difficult to detect.[5] Freedom House describes the 'liar's dividend' as a phenomenon in which convincing fake content makes even legitimate information suspect.[3] Iansiti and Lakhani warn that without transparency and oversight, generative AI could erode trust in institutions themselves.[6]

Understanding how trust is built, lost, or manipulated in the age of LLMs is essential to protecting the quality of public knowledge and maintaining healthy civic debate. The discussion proceeds in four sections: Section 1 reviews how researchers have defined trust in AI over time. Section 2 examines the role of LLMs in spreading misinformation and disinformation. Section 3 explores how LLMs create illusions of reasoning that affect user confidence. Section 4 synthesizes these findings and concludes with strategies to preserve trust in the context of rapid AI adoption.

Discussion

Defining Trust in Artificial Intelligence:

- ***Conceptual Definitions of Trust:***

Before examining how large language models affect trust in online information, it is necessary to clarify what 'trust in AI' means. Trust may mean believing the information is accurate, perceiving the source to be an expert, viewing the system as fair, or understanding how its outputs are produced. When studies use different definitions, comparisons across findings become difficult, and broader conclusions are limited.

Benk notes that many studies use inconsistent definitions of trust, and instead focus on isolated attributes such as reliance, transparency, or usability.[4] The 'Trustworthy AI' framework, developed by IBM, identifies common traits associated with trust, including accuracy, fairness, dependability, and transparency.[7] Research further shows that explainability, or understanding

DOI: 10.36838/IJHSR89.173

how AI arrived at the output, can support trust, but it does not guarantee reliance. Shin introduces the concept of 'causability', which emphasizes whether users can understand why an AI system produced a particular result, thereby increasing their confidence in using it.[8]

- ***Evolution of Trust in AI:***

The conceptual definitions of trust in AI have evolved. Earlier systems, like recommendation engines, were limited in scope and were often evaluated based on usefulness and consistency. In contrast, LLMs work across domains and generate content in human-like language. This increases their persuasive power while making errors difficult to detect. When the content resembles human writing, traditional credibility clues, like the author's identity or expertise, become less effective.

- ***Working Definition of Trust:***

Overall, trust in AI remains complicated and inconsistent in how it is defined or measured. Based on the reviewed literature, this review adopts a practical, working definition of trust in AI as a person's willingness to rely on an AI system's output because they believe it is accurate, fair, and produced through a process they can understand. This definition guides the analysis of how LLMs influence trust in online information.

Generative AI and the Spread of Misinformation and Disinformation:

- ***Distinguishing Misinformation and Disinformation:***

The rapid improvement of generative AI has altered the information landscape, amplifying both misinformation and disinformation. It is important to distinguish the two terms. Misinformation refers to incorrect and misleading information shared without an intent to deceive, while disinformation involves the deliberate generation of false information to mislead audiences. Disinformation poses greater risks when used strategically by bad actors.

- ***Social and Economic Impact of Generative AI:***

Freedom House (2023) reports that governments in at least forty-seven countries now deploy AI-generated content to influence online discourse, a number that has doubled over the past decade.[3] The MIT Technology Review similarly documents the use of generative AI to create political propaganda,[9] indicating that this technology is a force shaping world politics rather than remaining experimental.

One major consequence of this shift is the 'liar's dividend', a phenomenon in which the existence of convincing fake content causes users to doubt even authentic information. In such environments, bad actors do not need to produce highly persuasive lies; they only need to introduce enough doubt to make people distrust everything. A 2024 study by Google DeepMind and Jigsaw documented more than two hundred incidents of AI misuse in a single year, with deepfake media targeting public figures emerging as the most common form.[10]

The reach of AI-generated disinformation extends beyond politics. In the corporate world, scammers now use AI to impersonate executives with startling realism. The Wall Street Journal reported that more than 105,000 attacks using fake CEO voices or likenesses occurred in just one quarter of 2024, costing American companies over $200 million in losses.[11] RAND researchers warn that adversaries can now scale propaganda at unprecedented speed, not just in text but also through fabricated audio and video.[12]

These examples highlight that generative AI is accelerating changes to the information ecosystem. LLMs make misinformation cheaper, faster, and more believable. While propaganda and disinformation are not new phenomena, their scale and efficiency in the age of AI represent a concerning shift. As people begin to suspect even truthful content, the foundation of shared knowledge and democratic debate weakens.

Illusions of Reasoning and the Perceived Authority of LLMs:

- ***Fluency versus Reasoning:***

A defining feature of LLMs is their ability to generate fluent and coherent content even when it might be factually incorrect. Psychological research suggests that people are more likely to trust information that sounds confident and easy to understand. This is partly due to the fluency heuristic, where information that is easy to process feels more trustworthy. Huang and Chang (2023) note that LLMs replicate reasoning patterns without genuine comprehension, producing logical, structured language without the underlying understanding.[5] Users are also influenced by authority bias, interpreting structured and technical language as a sign of expertise.

- ***Chain-of-Thought and Perceived Transparency:***

Techniques such as chain-of-thought prompting deepen this perception by generating step-by-step explanations. Research shows that these explanations contain familiar rhetorical structures but may not reflect the internal probabilistic processes of the AI system. Studies from Arizona State University found that LLMs struggle when presented with unfamiliar tasks, despite sounding confident.[13] Anthropic similarly finds that chain-of-thought outputs often present post-hoc rationalizations rather than providing a transparent reasoning trace.[14] Unite.AI described this phenomenon as 'The mirage of AI reasoning'.[15]

- ***Empirical Findings of Misplaced Trust:***

A 2025 medRxiv preprint comparing ChatGPT, Gemini, and DeepSeek on primary care questions found significant variations in accuracy, while users reported the same level of trust regardless of performance.[16] LLMs, in essence, can be understood as performers of intelligence. Performance of reasoning rather than true reasoning strongly influences public trust.

Rebuilding Trust through Policy, Education, and Technical Measures:

- ***Governance and Accountability:***

Scholars and policymakers increasingly argue that strategies to rebuild trust in an AI-mediated world require standards for transparency, regulation, and platform accountability. Iansiti and Lakhani emphasize the importance of creating robust

governance frameworks to ensure responsible AI deployment.[6] Similar calls are echoed in healthcare. The American Medical Association published an eight-step AI Governance Toolkit to guide hospitals in managing ethical and technical risks.[17] These reinforce the need for proactive governance policies to prevent erosion of public trust.[18]

- ***Labeling and Human Oversight:***

Labeling AI-generated content and ensuring human oversight are additional measures. When platforms consistently disclose the true source of content creation or modification, users can assess credibility better. Labeling alone is not sufficient, but it may help users with contextual information. Andersen emphasizes that human oversight remains essential, particularly in public governance.[19]

- ***Education as a Long-Term Intervention:***

The World Economic Forum identifies AI literacy as a 'core competency' for students, and recommends training for critical evaluation of AI-generated output.[1] Education may be the most powerful long-term tool. Educational institutions are introducing digital literacy sooner. For example, Finland has integrated media literacy into its national education curriculum from an early age. Students are explicitly taught to evaluate sources, recognize misinformation, and question digital content. Common Sense Media released an updated K–8 digital literacy curriculum for young children about deepfakes and misinformation.[20,21] This long-term approach has been linked to higher public resilience against online misinformation. Research published in Scientific Reports shows that higher AI literacy correlates with more calibrated trust and improved evaluation of AI outputs.[22] These findings highlight the role of education leaders, teachers, parents, and policymakers in building the capacity for critical thinking in the next generation of internet users. Educational initiatives are central to helping users distinguish reliable and deceptive content.

- ***Technology Defenses and Their Limitations:***

Forensic detection tools developed by the U.S. National Institute of Standards and Technology (NIST) evaluate the authenticity of images, videos, and text.[23] Advances in multi-modal defenses, including voice pattern analysis and cross-referencing systems, offer potential for real-time analysis. While imperfect, technological advances make it harder for malicious actors to spread unchecked misinformation.

Conclusion

This paper has examined how large language models are fundamentally reshaping how trust operates in online environments. Section 1 showed that trust in AI is inconsistently defined and measured across studies, leaving major gaps in understanding. Section 2 demonstrated that generative AI is accelerating the spread of misinformation, contributing to the liar's dividend and weakening confidence in digital content. Section 3 revealed that users often mistake fluency for reasoning, which leads them to grant LLMs authority even when outputs are inaccurate. Section 4 reviewed possible solutions, including governance frameworks, content labeling, education, and technological detection tools.

The reviewed literature makes some points clear. Education improves users' ability to evaluate shortcomings of AI-generated content, encouraging better follow-up prompts. Accounting frameworks help reduce misuse and clarify responsibility. Together, education and governance emerge as the most effective and scalable approach in restoring calibrated trust, while labeling and detection tools play important secondary roles.

Several topics fall outside the scope of this literature review but are important for a holistic understanding and should be further investigated, including how different demographics and culture interpret AI-generated content, the corporate incentives surrounding perceived productivity gains and speed of delivery enabled by AI, distinctions between various categories of tools broadly referred to as AI, and how governments can implement transparency standards for better alignment with human values without stifling innovation. Clarifying what is well understood and what remains uncertain is essential for society to place appropriate trust in artificial intelligence for better outcomes.

Acknowledgments

The author would like to thank Dr. Siddharth Krishnan and Samuel Lefcourt from the University of North Carolina at Charlotte for their mentorship and guidance in reviewing and improving this manuscript.

References

1. Tanya Milberg, World Economic Forum. *Why AI literacy is now a core competency in education*. World Economic Forum, May 2025. https://www.weforum.org/stories/2025/05/why-ai-literacy-is-now-a-core-competency-in-education
2. Henrique, B., & Santos, E. Jr. Trust in artificial intelligence: Literature review and main path analysis. *Computers in Human Behavior: Artificial Humans*, *2*(1), 100043, 2024.
3. Allie Funk, Adrian Shahbaz, Kian Vesteinsson, *Freedom House. Freedom on the Net 2023: The repressive power of artificial intelligence*. Freedom House, 2023. https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence
4. Benk, T., *et al. Trust in artificial intelligence: Literature review and main path analysis*. Springer, 2024.
5. Huang, J., & Chang, K. Towards reasoning in large language models: A survey. *Findings of the Association for Computational Linguistics (ACL 2023)*, 2023. https://arxiv.org/abs/2212.10403
6. Iansiti, M., & Lakhani, K. R. *Competing in the age of AI: Strategy and leadership when algorithms and networks run the world*. Harvard Business Review Press, 2020.
7. Alice Gomstyn, Alexandra Jonker, Amanda McGrath, IBM. *What is trustworthy AI?* IBM Think. https://www.ibm.com/think/topics/trustworthy-ai
8. Shin, D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, 2021.
9. MIT Technology Review (Ryan-Mosley, T.). How generative AI is boosting the spread of disinformation and propaganda. *MIT Technology Review*, 2023. https://www.technologyreview.com/2023/07/27/generative-ai-propaganda-disinformation
10. DailyAI. DeepMind study exposes deepfakes as the leading form of AI misuse. *DailyAI*, June 2024. https://dailyai.com/2024/06/

DOI: 10.36838/IJHSR89.173

deepmind-study-exposes-deep-fakes-as-leading-form-of-ai-misuse
11. Angus Loten, Wall Street Journal. AI drives rise in CEO impersonator scams. *Wall Street Journal*, 2024. https://www.wsj.com/articles/ai-drives-rise-in-ceo-impersonator-scams-2bd675c4
12. Todd C. Helmus, RAND Corporation. *The rise of AI-enabled disinformation*. RAND Perspectives, 2024. https://www.rand.org/pubs/perspectives/PEA1043-1.html
13. Kyle Orland, Ars Technica. Researchers find LLMs are bad at logical inference, good at fluent nonsense. *Ars Technica*, August 2025. https://arstechnica.com/ai/2025/08/researchers-find-llms-are-bad-at-logical-inference-good-at-fluent-nonsense
14. Anthropic. Reasoning models don't say what they think. *Anthropic Research Blog*, May 2025. https://www.anthropic.com/research/reasoning-models-dont-say-think
15. Dr. Tehseen Zla, Unite.AI. The mirage of AI reasoning: Why chain-of-thought may not be what we think. *Unite.AI*, May 2025. https://www.unite.ai/the-mirage-of-ai-reasoning-why-chain-of-thought-may-not-be-what-we-think
16. medRxiv. Comparative accuracy of ChatGPT-o1, DeepSeek R1, and Gemini 2.0 in answering general primary care questions. *medRxiv Preprint*, 2025. https://doi.org/10.1101/2025.02.15.123456
17. American Medical Association. *Governance for augmented intelligence: Eight-step AI governance toolkit for health systems*. AMA STEPS Forward, 2025.
18. Dave Muoio, American Medical Association. AMA releases AI governance toolkit for health systems. *Fierce Healthcare*, August 18, 2025. https://www.fiercehealthcare.com/providers/ama-releases-ai-governance-toolkit-health-systems
19. Andersen, A. *AI and the quest for legitimacy in modern society: How the use of artificial intelligence in public governance impacts institutional legitimacy – The case of Norway. University of Bergen, 2023*.
20. Common Sense Media. Common Sense launches new digital literacy & well-being curriculum. *Common Sense Media*, August 18, 2025. https://www.commonsensemedia.org/press-releases/common-sense-media-launches-new-digital-literacy-well-being-curriculum-for-todays-classrooms
21. Common Sense Media. Deepfakes can be a crime: Teaching AI literacy can prevent it. *Common Sense Media*, May 15, 2025. https://www.commonsensemedia.org/kids-action/articles/deepfakes-can-be-a-crime-teaching-ai-literacy-can-prevent-it
22. Dongli Zhang, Tommy Tan Wijaya, Ying Wang, Mangy Su, Xinxin Li, Mia Wahyu Damayanti, Scientific Reports. Artificial intelligence literacy and trust in AI: Evidence from preservice teachers. *Scientific Reports*, *15*, 2025. https://doi.org/10.1038/s41598-025-99127-0
23. Haiying Guan, Andrew Zhang, Jim Horan, National Institute of Standards and Technology (NIST). *AI forensics evaluation report*. U.S. Department of Commerce, 2025. https://mig.nist.gov/MFC/Web/Papers/AFI2_ForensicsNIST_Final.pdf

Author

Ronav Lamba is a junior at Adlai E. Stevenson High School in Lincolnshire, a suburb of Chicago. He plans to major in engineering with a focus on robotics. His academic interests include mathematics, robotics, and artificial intelligence. He follows football, plays tennis, participates in fantasy football, coaches elementary basketball, and is a gym enthusiast.