AI & SOCIETY
https://doi.org/10.1007/s00146-025-02620-3

REVIEW

Generative AI and misinformation: a scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation

Seyeon Park[1] · **Xiaoli Nan**[1]

Received: 30 June 2025 / Accepted: 9 September 2025

Abstract
The rapid advancement of generative artificial intelligence (AI) has introduced both opportunities and challenges in the fight against misinformation. This scoping review synthesizes recent empirical studies to explore the dual role of generative AI—particularly large language models (LLMs)—in the generation, detection, mitigation, and impact of misinformation. Analyzing 24 empirical studies, our review suggests that LLMs can generate highly convincing misinformation, often exploiting cognitive biases and ideological leanings of the audiences, while also demonstrating the ability to detect false claims and enhance users' resistance to misinformation. Mitigation efforts show mixed results, with personalized corrections proving effective but safeguards inconsistently applied. Additionally, exposure to AI-generated misinformation was found to reduce trust and influence decision-making. This review underscores the need for standardized evaluation metrics, interdisciplinary collaboration, and stronger regulatory measures to ensure the responsible use of generative AI in the information ecosystem.

Keywords Generative AI · Misinformation · Fake news · Large language models · LLM

1 Introduction

In recent years, the proliferation of misinformation has emerged as one of the most pressing challenges facing contemporary society (Lewandowsky 2023; Swire-Thompson and Lazer 2020). While definitions vary, misinformation is broadly understood as false or misleading information shared without malicious intent (Ireton and Posetti 2018), and, more broadly, as an umbrella term that includes all forms of false or misleading information, including fake news, conspiracy theories, pseudoscience, and propaganda (Ecker et al. 2022). In the health and science context, misinformation typically refers to content that contradicts expert consensus (Nan et al. 2023; Swire-Thompson and Lazer 2020; Vraga and Bode 2020). Its societal impact is particularly acute in domains such as health, where it has been linked to vaccine hesitancy, resistance to public health measures, and confusion during crises like the COVID-19 pandemic (Nan et al. 2022a).

Compounding this challenge is the rise of generative artificial intelligence (AI), a rapidly evolving class of technologies capable of producing coherent, human-like text, images, audios, and videos. Powered by large language models (LLMs) such as OpenAI's GPT series and Google's Bard, generative AI systems now function not merely as information retrieval tools but as autonomous content creators (Cascella et al. 2023). Their fluency, scale, and adaptability offer unprecedented opportunities—and risks—for how information is generated, consumed, and trusted. While generative AI holds promise for combating misinformation through tools for detection and correction, it also introduces new vectors for harm. One critical concern is the phenomenon of AI "hallucinations"—instances where LLMs confidently produce factually inaccurate responses (Bandara 2024). Such content, when delivered with persuasive language and without disclaimers, can mislead users who interpret AI outputs as authoritative or objective, a tendency rooted in the "machine heuristic" (Sundar 2008). Moreover, generative AI may be intentionally exploited by bad actors to fabricate convincing disinformation, synthetic media, or counterfeit scientific reports (Kim et al. 2024).

✉ Seyeon Park
spark143@umd.edu

Xiaoli Nan
nan@umd.edu

[1] University of Maryland, College Park, College Park, USA

Published online: 30 September 2025

Despite growing interest in the intersection of AI and misinformation, research remains fragmented, with studies focusing on disparate applications, models, and contexts. Existing reviews often emphasize health or political misinformation broadly, without systematically examining the unique implications introduced by generative AI (Pérez-Escolar et al. 2023). To address this gap, this scoping review synthesizes recent empirical studies on the role of generative AI in the generation, detection, mitigation, and impact of misinformation. By charting how generative AI such as LLMs both exacerbate and alleviate misinformation challenges, this review provides a comprehensive overview of the technological, psychological, and ethical dimensions of AI-mediated communication. In doing so, it offers a foundational understanding for scholars, practitioners, and policymakers seeking to navigate the double-edged nature of generative AI in today's complex information ecosystem.

2 Conceptual background

2.1 An overview of misinformation research

Over the past decade, misinformation has become a critical concern in communication research, particularly as digital platforms facilitate the rapid dissemination of false or misleading content (Del Vicario et al. 2016; Muhammed and Mathew 2022; Nan et al. 2022b; Suarez-Lledo and Alvarez-Galvez 2021; Wang et al. 2019). Misinformation poses significant challenges across multiple domains, threatening democratic processes (e.g., Lewandowsky 2023; Walter and Tukachinsky 2020), increasing false beliefs (Schmid et al. 2023), eroding public trust (e.g., Di Domenico and Ding 2023; Ecker et al. 2024; Ognyanova et al. 2020), and undermining informed health decision-making such as vaccination (e.g.,Nan et al. 2023; Neely et al. 2022; Swire-Thompson and Lazer 2020; Zimmerman et al. 2023). In response, scholars have examined a range of misinformation interventions in recent years. One widely studied approach involves debunking, which provides corrections after exposure to misinformation, can reduce belief in misinformation, but its efficacy depends on cognitive elaboration and worldview alignment (Chan et al. 2017; Walter and Tukachinsky 2020). Psychological inoculation or prebunking, which offers preemptive warnings and refutations, has also shown promise in building resistance to misinformation (Van der Linden et al. 2017; Traberg et al. 2022).

While misinformation research has advanced our understanding of misinformation and its mitigation and impact, the rapid evolution of technology has introduced new challenges. The emergence of generative AI has transformed the misinformation landscape, enabling the creation of highly realistic fake content at an unprecedented scale. Unlike previous forms of misinformation that relied on human manipulation, AI systems can now generate convincing text, images, and videos that blur the line between fact and fiction. This shift raises critical concerns about the generation and psychological impact of AI-generated misinformation and about the potential to leverage generative AI for detecting and mitigating false content. Further research into misinformation within the context of generative AI is crucial for understanding the evolving dynamics of AI-driven misinformation.

2.2 Generative AI and misinformation

Generative AI refers to the type of AI capable of generating new, creative content such as text, images, audio, or videos from training data (Aydin and Karaarslan 2023). Unlike earlier forms of AI, generative AI systems can autonomously produce human-like outputs from input prompts, powered by LLMs, deep learning models pre-trained on a vast corpus of textual data to predict and generate coherent language (Cascella et al. 2023). These systems mark not just a technical evolution but a paradigmatic shift in the communication landscape. Unlike traditional search engines, which require users to navigate multiple sources to locate information, LLMs offer direct, synthesized responses tailored to user prompts (Zhou and Li 2024). Their persuasive, context-sensitive fluency positions them as influential agents in shaping public discourse and knowledge production, raising urgent questions about their role in amplifying or mitigating misinformation.

A major concern is hallucinations, where LLMs fabricate plausible but false information (Bandara 2024; Kamel 2024; Monteith et al. 2024; Sun et al. 2024). Hallucinations often arise from training data gaps and probabilistic predictions (Athaluri et al. 2023; Kamel 2024). Compounding the risk, users often perceive AI outputs as objective and credible, reflecting the "machine heuristic" (Monteith et al. 2024; Sundar 2008). This tendency makes hallucinated content, especially concerning, as it may be uncritically accepted as fact. Intentional misuse of generative AI presents a distinct and equally concerning challenge. The term "deepfake," once narrowly associated with face-swapping technology, now encompasses a wide range of AI-generated media, including voice imitation and synthetic videos (Tolosana et al. 2020; Verdoliva 2020). These highly realistic fabrications have been increasingly used in disinformation campaigns, particularly targeting political figures (Westerlund 2019), raising concerns about their impact on public perception and trust (Dobber et al. 2021; Vaccari and Chadwick 2020; Verma 2024). The ease of creating convincing deepfakes, removing the barrier of technical expertise, further amplifies the threat (Romero Moreno 2024).

At the same time, generative AI offers opportunities for misinformation detection and mitigation. LLMs possess strong reasoning ability (Huang and Chang 2022) and can verify factual errors based on extensive pretrained data (Augenstein et al. 2024; Kuznetsova et al. 2025). For example, Kuznetsova et al. (2025) conducted a comparative study to evaluate the effectiveness of ChatGPT and Bing Chat in assessing the veracity of political information in multiple languages, with ChatGPT achieving over 81% accuracy in classifying conspiracy theories. Recent experiments have also shown that conversational generative AI fact-checking can yield positive outcomes (e.g., Costello et al. 2024; Lu 2025).

Despite these emerging insights, there is a lack of synthesized evidence that comprehensively maps the evolving functions and implications of generative AI in the context of misinformation. Thus, this scoping review seeks to systematically examine the role of generative AI in the generation, detection, mitigation, and impact of misinformation. By synthesizing recent empirical studies, the review offers a comprehensive overview of the ways generative AI both exacerbates and addresses pressing misinformation challenges, highlighting the risks it poses as well as the opportunities it affords.

3 Method

This scoping review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) guidelines (Tricco et al. 2018) to ensure a structured and transparent methodology. Figure 1 presents a PRISMA flow diagram, illustrating the step-by-step process of searching, screening, and including studies in the final review.

3.1 Search strategy

Given the interdisciplinary nature of the research areas, we used Google Scholar for literature search. Google Scholar

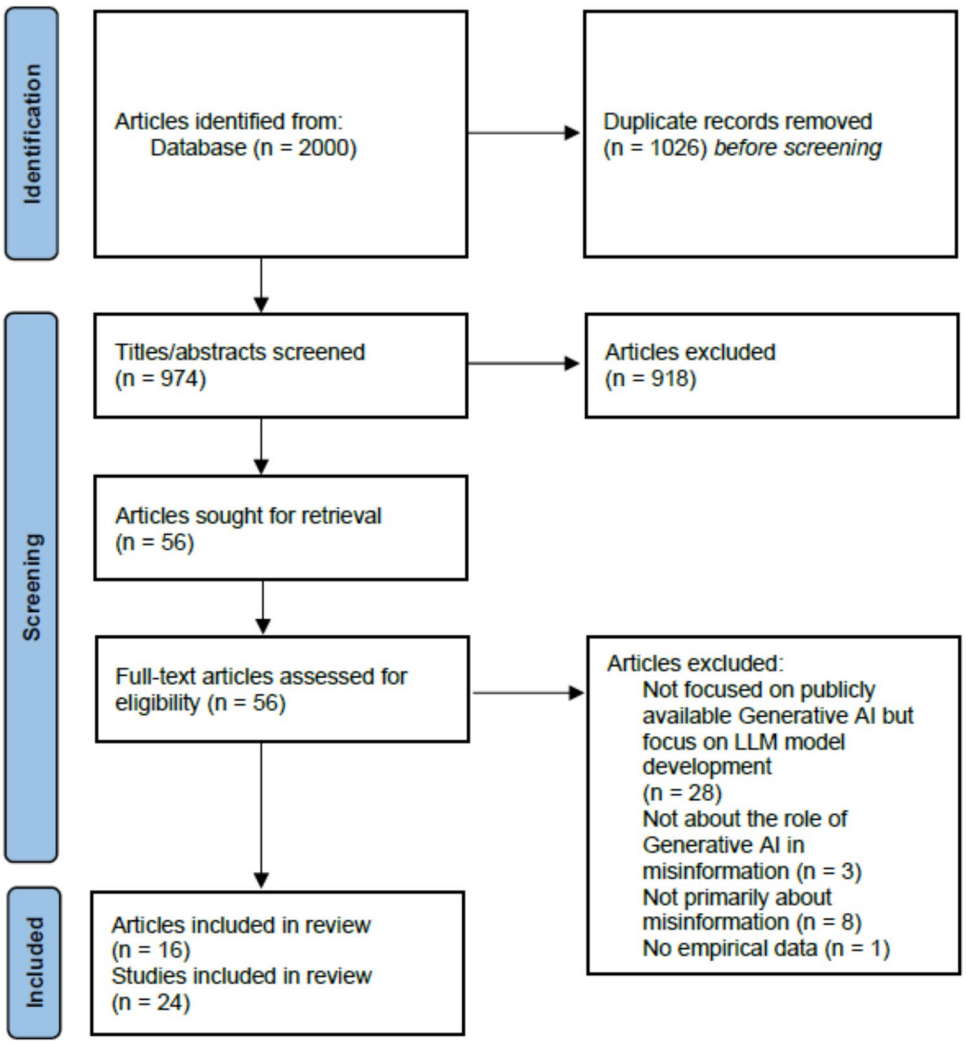

Fig. 1 PRISMA flow diagram

offers broad and rapid indexing across a wide range of disciplines, including computer science, communication, public health, and the social sciences—fields highly relevant to our review. This breadth allows for the inclusion of the most recent and often-cited studies, which is essential in fast-evolving domains like generative AI and misinformation. Our search covered articles published from January 2018 to September 19, 2024, aligning with the introduction of the first-generation GPT models. To capture relevant studies, we combined terms related to generative AI (e.g., "large language model," "LLM," "GPT," "ChatGPT," "generative AI") and misinformation (e.g., "misinformation," "disinformation," "fake news," "propaganda"), using Boolean AND operators. For each search query, the first 100 results were retained for further screening, following the protocols of previous research (Noar et al. 2016).

3.2 Screening process

We screened each article identified in the literature search and the study or studies reported in each article. The selection process was conducted in two stages: (1) title and abstract screening and (2) full-text review. One author conducted the initial title and abstract screening. Then the two authors independently conducted full-text screening. Studies were screened based on predefined inclusion and exclusion criteria. Both title/abstract and full text screening relied on the following inclusion/exclusion criteria. To be included, studies must be empirical, meaning they report original data rather than theoretical or conceptual discussions. The studies also had to examine at least one of the following areas: the capacity of generative AI such as LLMs to create misinformation, their capacity to detect or correct misinformation, or the effects of AI-generated misinformation on audiences. Additionally, studies addressing one or more of the four topics must focus on publicly available generative AI models, including LLMs such as ChatGPT series or Google Bard, rather than highly fine-tuned models or software that are built on current AI models. Furthermore, to ensure the methodological rigor of included studies, only peer-reviewed journal articles or conference proceedings were retained, and both quantitative and qualitative research approaches were considered. Finally, all included studies had to be published in English.

3.3 Coding process

To ensure a systematic and reliable approach to data extraction, a codebook was developed through iterative pilot coding and discussion. This process aimed to refine the variables, definitions, and coding criteria, ensuring that all coders applied the codes consistently across studies. Both authors independently coded a subset of the dataset using the initial version of the codebook. After this initial round of coding, discrepancies were identified and reviewed collaboratively. These discussions led to refinements in the codebook's structure to improve alignment between coders. Both authors then coded all studies independently based on the refined codebook.

To assess intercoder reliability, we calculated agreement metrics across all key variables. For most variables, including Country of Authors, LLM Type, Context, Key Results, and Conclusions/Implications, coders demonstrated 100% agreement ($k = 1.00$). Minor wording differences were resolved via consensus discussions, with no substantive disagreements requiring adjudication. For the variable Study Type, the initial agreement was 81.25% with a Cohen's kappa (k) ranging from 0.60 to 1.00, with an average kappa of 0.822, indicating substantial agreement. For Evaluation Method, agreement was 75% with $k = 0.50$, reflecting moderate agreement. To address discrepancies, the coders jointly reviewed and discussed all mismatched cases until consensus was reached. All remaining discrepancies were resolved through consensus-based adjudication to ensure consistency in the final dataset. The complete codebook, including variable names, definitions, coding categories, and instructions, is presented in Supplementary Table S1. Each publication was documented with its authors, year of publication, publication outlet, and country of origin. Studies were then categorized based on the primary research focus into one or more of the following categories: (1) AI generation of misinformation, (2) AI detection of misinformation, (3) AI-driven mitigation strategies to counter misinformation, (4) impact of AI-generated misinformation on users, and (5) other. The evaluation methods employed in each study to assess AI capacity were classified into three categories: (1) human evaluation, where experts or raters directly assessed AI outputs; (2) human experiments, where studies examined the effects of user interactions with generative AI; and (3) other methods, such as comparisons with fact-checked datasets or external data to measure accuracy and reliability of AI outputs. The type of generative AI analyzed and the context of misinformation addressed in each study, such as health, politics, or other relevant areas, were recorded as well.

4 Results

4.1 Study selection

A total of 2000 records were identified through Google Scholar. Before screening, 1026 duplicate records were identified and removed, leaving 974 unique articles for title and abstract screening. At this stage, each article underwent an initial review and its relevance based on the predefined inclusion and exclusion criteria was assessed. Following

the title and abstract screening, 918 records were excluded, resulting in 56 full-text articles being evaluated for eligibility. During this stage, 40 studies were excluded. The primary reasons for exclusion included a lack of focus on publicly available generative AI models (28 studies), a lack of direct examination of generative AI's role in misinformation (3 studies), and a primary focus on propaganda without explicitly addressing misinformation (8 studies). Additionally, one study was excluded for lacking empirical data. After applying the inclusion/exclusion criteria, 16 articles were retained for data extraction and analysis. These articles included 24 empirical studies, each examining various aspects of generative AI and misinformation, including detection, generation, mitigation strategies, and impact on audiences.

4.2 Study characteristics

Table 1 presents the characteristics of the 16 articles included in this scoping review. The 16 articles included in the review were published between 2023 and 2024, with nine articles published in 2023 and seven in 2024. Figure 2 presents the geographic distribution of studies by country. Based on the first author's affiliation, most articles originated from North America and Europe. Across 16 articles, the United States (n = 4) had the highest representation, followed by Italy (n = 2), Switzerland (n = 2), and Australia (n = 2). Other articles included authors from South Africa, Spain, Poland, China, India, and South Korea. Several articles involved international collaboration (n = 7), with co-authors representing diverse institutional affiliations.

Figure 3 illustrates the distribution of study characteristics in evaluation methods, study focus, and the types of generatiev AI models. The 24 studies examined generation (n = 10), detection (n = 6), mitigation (n = 4), and impact (n = 7), with some studies spanning multiple categories. Across the 24 studies, human experiments were employed in ten studies (e.g., Gabriel et al. 2024; Kim et al. 2023), where participants evaluated AI-generated content or were exposed to AI message interventions. Human evaluation was also used to assess AI outputs (n = 8), with expert raters comparing AI-generated responses to factual information or established benchmarks (e.g., Deiana et al. 2023; Menz et al. 2024). In two studies, human evaluation was combined with computational methods (Sparks et al. 2024; Wang et al. 2023). Others relied solely on computational comparisons with fact-checked datasets (Kumar et al. 2024, Study1-2; Santangeli et al. 2024; Węcel et al. 2023). Among the generative AI models, ChatGPT was the most commonly studied (n = 27), followed by Google Bard (n = 4), Perplexity AI (n = 2), and Microsoft Bing Chat (n = 2), with some studies spanning multiple models (e.g., Garbarino and Bragazzi 2024; Kumar et al. 2024). Regarding the study topic, health misinformation was the most frequently studied (n = 6), including vaccination (Deiana et al. 2023), abortion (McMahon and McMahon 2024), COVID-19 (Wang et al. 2023), and other health topics (Menz et al. 2024; Garbarino and Bragazzi 2024; Sparks et al. 2024). Political and conspiracy misinformation was another key focus (n = 5), including false news and conspiracy theories (Gabriel et al. 2024), Russian-Ukraine war (Makhortykh et al. 2024), and South African conspiracy theories (Senekal and Brokensha 2023). Other domains included science (Spitale et al. 2023), wildlife (Santangeli et al. 2024), and travel (Kim et al. 2023, Studies 1–5). Some studies examined multiple domains, including politics, health, economics, science, religion, ethical dilemmas, and pseudoscience (Kim et al. 2024; Kumar et al. 2024; Węcel et al. 2023).

4.3 Main findings

4.3.1 *Misinformation generation* via *generative AI*

A total of ten studies (Gabriel et al. 2024, Study 3; Kim et al. 2024; Kumar et al. 2024, Study 2; Makhortykh et al. 2024; McIntosh et al. 2023; McMahon and McMahon 2024; Menz et al. 2024; Senekal and Brokensha 2023; Sparks et al. 2024; Wang et al. 2023) examined the capacity of LLMs to generate misinformation across political, health, and general misinformation domains. Several studies highlighted the deceptiveness of AI-generated misinformation. Gabriel et al. (2024, Study 3) examined ChatGPT-4.0's capacity to generate personalized misinformation. They prompted ChatGPT-4.0 to create false news headlines based on common conspiracy theories, such as vaccine and flat Earth narratives, and tailored the language to participants' demographic profiles such as age, education, political ideology, race, and gender. They found that ChatGPT-4.0-generated personalized misinformation was difficult to detect, especially when aligned with participants' demographics.

Similarly, Kim et al. (2024) showed how generative AI can convincingly manipulate numerical, textual, and visual data. ChatGPT-4.0 was able to manipulate numerical datasets to achieve statistically or clinically significant outcomes, adjust interdependent economic indicators based on fabricated inputs, and alter the sentiment and content of interview transcripts while maintaining stylistic consistency. Adobe Firefly, a visual generative model, successfully inserted visual fabrications, such as inserting liquid water into a Mars rover image. The study also identified methods for bypassing built-in safeguards through prompt engineering and iterative refinement, underscoring ethical vulnerabilities in data-driven misinformation. Kumar et al. (2024, Study 2) examined the linguistic characteristics of LLM-generated misinformation, focusing on how ChatGPT-3.5 distorts news content. Compared to authentic news and human-generated misinformation, LLM-generated misinformation exhibited

Table 1 Study characteristics

#	Author (year)	Study topic	Country	Study focus	Evaluation method	Generative AI model
1	Deiana et al. (2023)	Health	Italy, Switzerland	Detection	Human evaluation	ChatGPT-3.5, ChatGPT-4.0
2	Gabriel et al. (2024) Study 1	Politics, conspiracy theories	USA	Mitigation	Human experiment	ChatGPT-4.0
3	Gabriel et al. (2024) Study 2	Politics, conspiracy theories	USA	Mitigation	Human experiment	ChatGPT-4.0
4	Gabriel et al. (2024) Study 3	Politics, conspiracy theories	USA	Generation	Human experiment	ChatGPT-4.0
5	Garbarino and Bragazzi (2024)	Health	Italy, Canada	Detection	Human evaluation	ChatGPT-4.0, Google Bard
6	Kim et al. (2023) study 1	Travel	Korea, New Zealand, China, Japan, USA	Impact	Human experiment	ChatGPT-3.0
7	Kim et al. (2023) study 2A	Travel	Korea, New Zealand, China, Japan, USA	Impact	Human experiment	ChatGPT-3.5
8	Kim et al. (2023) study 2B	Travel	Korea, New Zealand, China, Japan, USA	Impact	Human experiment	ChatGPT-3.5
9	Kim et al. (2023) study 3	Travel	Korea, New Zealand, China, Japan, USA	Impact	Human experiment	ChatGPT-3.0
10	Kim et al. (2023) study 4	Travel	Korea, New Zealand, China, Japan, USA	Impact	Human experiment	ChatGPT-3.0
11	Kim et al. (2023) study 5	Travel	Korea, New Zealand, China, Japan, USA	Impact	Human experiment	ChatGPT-3.0
12	Kim et al. (2024)	Multiple domains	USA	Generation	Human evaluation	Adobe Firefly (visual gen-AI), ChatGPT-4.0
13	Kumar et al. (2024) study 1	Multiple domains	India, Austria	Detection	Other	ChatGPT-3.5, FLAN-T5, GPT-BLOOM, GPT-Neo
14	Kumar et al. (2024) study 2	Multiple domains	India, Austria	Generation	Other	ChatGPT-3.5
15	Makhortykh et al. (2024)	Russia-Ukraine War	Switzerland, Germany	Generation, Mitigation	Human evaluation	Google Bard, Microsoft Bing, Perplexity
16	McIntosh et al. (2023)	Multiple domains	Australia	Generation	Human evaluation	ChatGPT-3.5, ChatGPT-4.0, Google Bard, Perplexity, TruthGPT
17	McMahon and McMahon (2024)	Health	USA	Generation	Human evaluation	ChatGPT-3.5
18	Menz et al. (2024)	Health	Australia, USA, Canada, UK	Generation, Mitigation	Human evaluation	ChatGPT-4.0, Copilot, Google Bard, HuggingChat, Poe
19	Santangeli et al. (2024)	Wildlife	Spain, South Africa, Italy, Argentina	Detection	Other	ChatGPT-3.5, ChatGPT-4.0, Microsoft Bing
20	Senekal and Brokensha (2023)	Conspiracy theories	South Africa	Detection, Generation	Human evaluation	ChatGPT (version not specified)
21	Sparks et al. (2024)	Health	USA	Generation	Human evaluation, Other	ChatGPT-3.5
22	Spitale et al. (2023)	Science	Switzerland	Impact	Human experiment	ChatGPT-3.0
23	Wang et al. (2023)	Health	China	Generation	Human evaluation, Other	ChatGPT-3.5, ChatGPT-4.0
24	Węcel et al. (2023)	Multiple domains	Poland	Detection	Other	ChatGPT-3.5

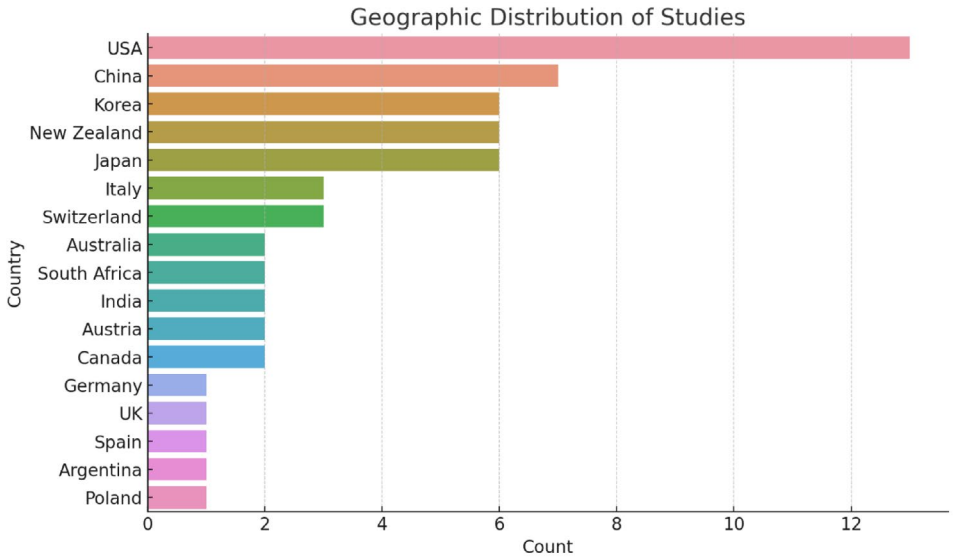

Fig. 2 Geographic distribution of studies

higher concreteness, lower abstractness, and reduced named entity density (i.e., fewer references to specific people, organizations, or places relative to word count). These patterns were consistent across different prompting strategies and source texts.

Several studies highlighted the risks of LLM-generated misinformation in high-stakes contexts, revealing inconsistent reliability and a tendency to reproduce or amplify false narratives. Makhortykh et al. (2024) investigated the potential of generative AI to generate and reinforce geopolitical misinformation, with a focus on Russian disinformation narratives concerning the war in Ukraine. The study found that more than 25% of responses failed to meet expert-verified standards, often reproducing false or misleading claims without adequate disclaimers or corrective context across three widely used generative AI systems (Google Bard, Bing Chat, and Perplexity AI). In the health domain, McMahon and McMahon (2024) identified significant misinformation from ChatGPT regarding self-managed abortion, notably overstating risks and contradicting established medical guidelines. Sparks et al. (2024) reported that although ChatGPT responses about orthopedic conditions were generally accurate, they lacked detailed responses regarding orthopedic conditions. Wang et al. (2023) showed that ChatGPT-3.5 and 4.0 generated moderately accurate COVID-19 content, but notable gaps remained compared to official sources like the World Health Organization (WHO).

McIntosh et al. (2023) developed a Culturally Sensitive Test to evaluate five LLMs—ChatGPT-3.5, ChatGPT-4.0, Google Bard, Perplexity AI, and TruthGPT—to generate hallucinated or false content across 70 prompts spanning seven contentious domains, including politics, ethics, pseudoscience, health, and social norms. The study found that LLMs were significantly more likely to generate hallucinated or incoherent responses in politically and culturally sensitive domains compared to health or scientific topics. This context-dependent pattern of output suggests that misinformation generation is more likely to occur in areas shaped by cultural subjectivity or moral disagreement, where language models struggle to anchor their responses to verifiable facts.

Senekal and Brokensha (2023) investigated ChatGPT's potential to generate misinformation by examining its responses to ten South African conspiracy theories. While ChatGPT did not actively propagate South African conspiracy theories and generally produced accurate information, it reproduced one specific misinformation by falsely claiming that Hamilton Naki assisted in the world's first human heart transplant. This false narrative—originally published by sources such as the *New York Times* and the *Lancet* but later corrected—was likely present in the model's training data. Though unintentional, this response illustrates how LLMs can generate misinformation when trained on content that includes uncorrected or outdated narratives from authoritative sources. The study also observed a left-leaning political bias in ChatGPT's responses to politically sensitive prompts, suggesting that LLM-generated content may not only reflect factual inaccuracies but also ideological imbalances.

Menz et al. (2024) evaluated the ability of five LLMs to generate health-related misinformation. The study found that ChatGPT, Google Bard, and HuggingChat consistently generated cancer misinformation blogs, totaling over 40,000 words across 113 unique outputs, without requiring jailbreaking. These outputs included fabricated academic citations, clinician and patient testimonials, and demographic targeting, demonstrating the LLMs' capacity for scalable, tailored misinformation generation. On the other hand, Claude 2.0 and Copilot were effective in resisting misinformation prompts, even under jailbreaking attempts. However,

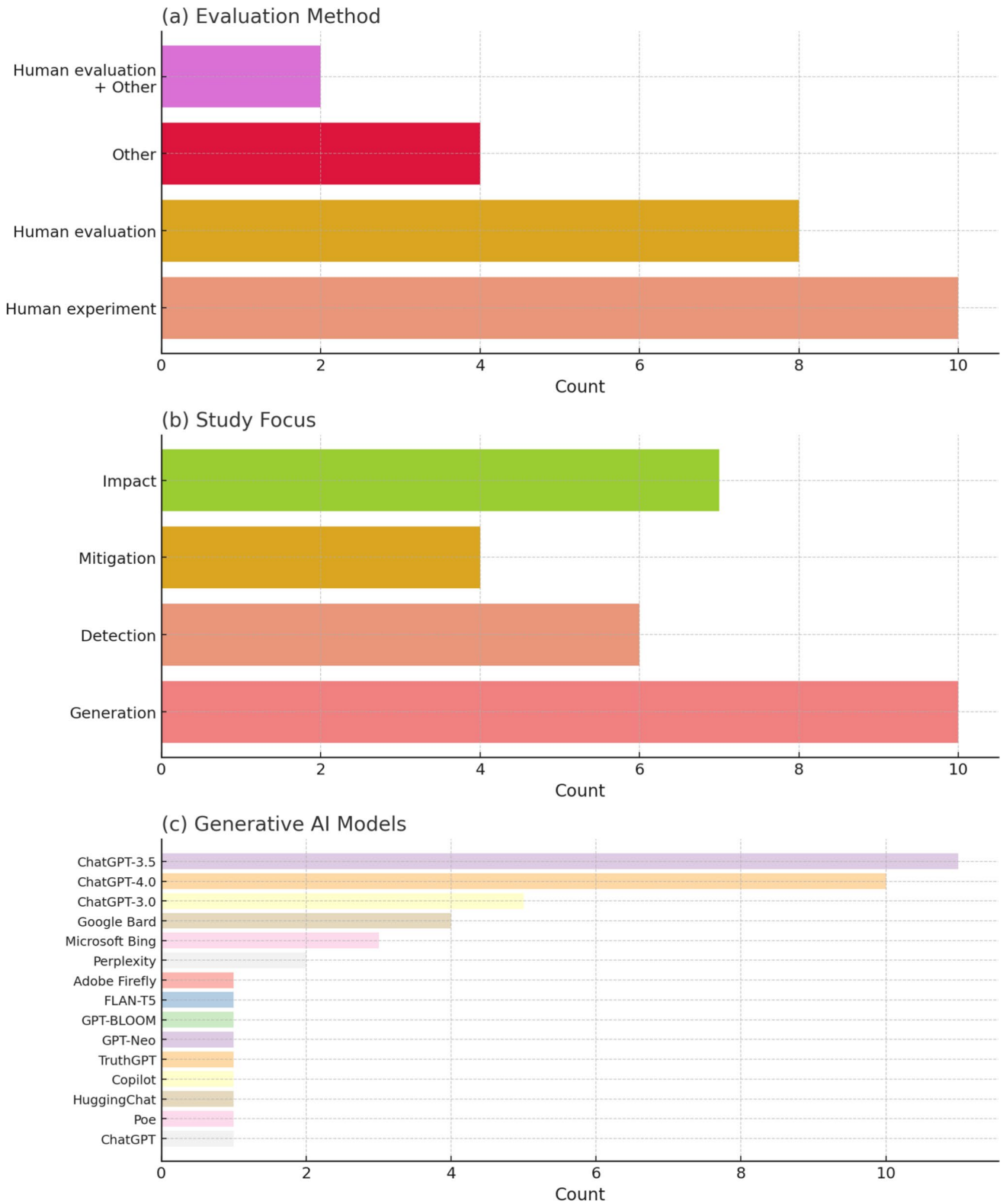

Fig. 3 Overview of study characteristics. Distribution of studies by **a** evaluation method, **b** study focus, and **c** generative AI models in studies

Copilot later generated misinformation in 12-week follow-up tests without the need for jailbreak prompts.

4.3.2 *Misinformation detection* via *generative AI*

A total of six studies (Deiana et al. 2023; Garbarino and Bragazzi 2024; Kumar et al. 2024, Study 1; Santangeli et al. 2024; Senekal and Brokensha 2023; Węcel et al. 2023) investigated the ability of LLMs to detect misinformation across diverse topics, including health and conspiracy theories. These studies highlighted both strengths and notable limitations in AI-driven misinformation detection.

Deiana et al. (2023) evaluated the reliability of ChatGPT-3.5 and GPT-4.0 in detecting and addressing misinformation about vaccines, based on the World Health Organization's 11 most common vaccine myths. Both versions generally provided accurate, comprehensive, and conversational responses, with GPT-4.0 consistently outperforming GPT-3.5 in terms of correctness, clarity, and exhaustiveness. Notably, the average accuracy of responses improved when questions were presented in sequence, indicating greater performance when contextual information was preserved. Despite these promising results, the study underscored the potential risks of relying on AI-generated health information without expert oversight, especially for non-expert users who may struggle to detect subtle inaccuracies or contextual gaps. Similarly, Garbarino and Bragazzi (2024) compared the accuracy of ChatGPT-4.0 and Google Bard in assessing sleep-related misinformation. The study found a moderately strong agreement between ChatGPT-4.0 and Google Bard evaluations of sleep-related misinformation and expert assessments. However, the level of alignment varied based on whether claims were assessed for factual falseness or their public health significance, underscoring context-dependent reliability. While Google Bard slightly outperformed ChatGPT-4.0 in accurately identifying false claims, ChatGPT-4.0 demonstrated stronger alignment with expert assessments, particularly in evaluating both the factual falseness and the public health significance of the myths.

Kumar et al. (2024) focused on evaluating multiple LLMs—GPT-3.5, GPT-Neo, FLAN-T5, and BLOOM in detecting misinformation across six benchmark datasets in Study 1. A key finding was that zero-shot prompting, where a model is asked to classify misinformation without being shown any examples, generally outperformed few-shot prompting, which involves providing a small number of labeled examples beforehand. This counterintuitive result was attributed to the noise and inconsistency that few-shot examples can introduce, which may confuse the model rather than improve its ability to learn patterns. Although few-shot learning is typically expected to enhance model performance by offering task-specific context, in this case, the examples appeared to introduce variability and ambiguity that interfered with the models' reasoning. The study also examined the effect of including sentiment and emotional cues—such as labeling news content with emotions like "anger" or "joy" in datasets, finding that incorporating these features reduced detection accuracy, particularly in zero-shot settings.

Santangeli et al. (2024) analyzed LLMs' ability to detect fake and sensationalized wildlife news. The study identified a positive correlation between AI-generated likelihood risk scores and actual risk levels, indicating that LLMs can approximate real-world data when evaluating animal threats. However, performance varied by context. Specifically, LLMs exhibited weaker detection capabilities for misinformation related to human threats but performed better when assessing misinformation about livestock attacks. Senekal and Brokensha (2023) evaluated ChatGPT's capacity in detecting misinformation by analyzing its responses to ten South African conspiracy theories. To conduct this evaluation, they first converted the conspiracy theories into yes or no questions to serve as prompts, such as: "Did the CIA develop the Human Immunodeficiency Virus (HIV) to kill Africans?". While ChatGPT generally provided accurate "no" responses to most conspiracy theories, correctly rejecting the false claims, it failed to detect one conspiracy theory. This was likely due to the presence of the false claim in widely trusted sources such as the *New York Times*, the *Lancet*, and the *British Medical Journal*, which are representative of the types of mainstream sources that inform ChatGPT's training data.

Węcel et al. (2023) evaluated ChatGPT's ability to detect fake news and found significant variability in ChatGPT's accuracy in detecting fake news by comparing its classifications to human fact-checkers' judgments across multiple claims. To test the consistency and reliability of ChatGPT's outputs, the researchers employed six different prompt formats, each representing a distinct but semantically equivalent way of asking the model to evaluate a claim. These prompt formats varied in tone and structure from direct questions like "Is this claim true or false?" to more cautious or evaluative instructions such as "Evaluate the following claim and indicate whether there is sufficient evidence to support it." Despite all prompts aiming to elicit the same type of response, the study found that accuracy to verify claims varied significantly based on prompt wording, with some formats producing more assertive or more cautious responses. Overall, ChatGPT's accuracy remained low, only slightly better than random guessing, and agreement with human fact-checkers ranged from slight to fair. Performance was higher for English-language claims than Polish ones, and claims published before ChatGPT's 2021 training cutoff did not yield improved results.

4.3.3 Mitigation of AI-generated misinformation

Four studies, published across three articles (Gabriel et al. 2024, Study 1, Study 2; Makhortykh et al. 2024; Menz et al. 2024), empirically examined the potential of generative AI to mitigate the impact of AI-generated misinformation. Gabriel et al. (2024) present two key studies investigating the role of generative AI in combating misinformation. The first study, conducted in a simulated social media environment, tests the effectiveness of five non-personalized interventions in helping users identify false content: (1) a simple label indicating whether a claim is true or false, (2) a methodology-based explanation stating that an AI model verified the claim, (3) a similar methodology-based explanation attributing verification to human fact-checkers, (4) a reaction-frame explanation that describes the intent behind the claim, and (5) a GPT-4-generated explanation providing a short rationale for the claim's veracity. Findings show that explanation-based interventions outperform label-only approaches, with GPT-4 explanations yielding the highest accuracy improvement (up to 47.6%) and the most significant reduction in misinformation sharing. The second study examines the effectiveness of personalized interventions, where GPT-4-generated explanations are tailored to users' demographics (e.g., education level, political ideology, age, and gender). Results indicate that personalized explanations are rated as more helpful than generic ones, particularly when well-aligned with users' attributes. However, misaligned personalizations reduce effectiveness. Specifically, when personalization was misaligned—such as an explanation designed for a liberal-leaning individual being shown to a conservative user—the intervention's perceived credibility decreased, sometimes making users more resistant to accepting the correction.

Furthermore, Makhortykh et al. (2024) examine how LLM-powered chatbots handle Russian disinformation narratives about the war in Ukraine. Using an AI audit methodology, the researchers manually tested 28 prompts related to Kremlin-sponsored disinformation by submitting each prompt four times to each chatbot. They then assessed chatbot outputs against expert baselines to evaluate accuracy, inclusion of disclaimers, and consistency. The findings reveal that more than a quarter of chatbot responses propagate false or misleading information. Less than half of the responses acknowledged the Russian perspective on war-related issues, and when they did, 7 to 40% failed to debunk Kremlin disinformation. Additionally, the study highlights a concerning level of inconsistency, where identical prompts yielded dramatically different chatbot responses, potentially exposing users to contradictory information. This variation, attributed to the stochastic nature of LLMs, raises concerns about their reliability in combating misinformation in politically sensitive contexts.

Finally, Menz et al. (2024) evaluate the effectiveness of safeguards in preventing LLMs from generating health disinformation. The researchers tested four widely used LLMs by prompting them to generate misinformation about two controversial health topics: the claim that sunscreen causes skin cancer and the assertion that the alkaline diet cures cancer. They assessed whether the models could be jailbroken to bypass safeguards and analyzed the transparency of AI developers regarding risk mitigation. The findings revealed that while Claude 2 consistently refused to generate disinformation even with jailbreaking attempts, other tested LLMs freely generated 113 unique health disinformation blogs. A follow-up evaluation 12 weeks later found that safeguards had weakened, with GPT-4 (via Copilot) beginning to generate disinformation despite initially blocking such content. These results underscore the insufficiency of current mitigation measures against the misuse of LLMs in spreading harmful health misinformation.

4.3.4 Impact of AI-generated misinformation

A total of seven studies reported in two articles (Kim et al. 2023; Spitale et al. 2023) empirically tested the impact of AI-generated misinformation. Across six experiments, Kim et al. (2023) examined the impact of incorrect information from ChatGPT on travelers' acceptance of AI-generated recommendations. When ChatGPT provided incorrect information, visit intentions declined, even with explicit error reminders (Study 1). This effect persisted across different participant pools and without explicit reminders (Study 2A, 2B). The impact of misinformation was reduced when it appeared earlier in a list (Study 3). Prior exposure to incorrect information negatively influenced subsequent decision-making, particularly when within the same domain (Study 4), though this effect diminished when participants focused on selecting an option rather than evaluating accuracy (Study 5). Collectively, these studies highlight the potential of AI-generated misinformation to mislead consumer decisions as well as conditions that alleviate and exacerbate the harmful effects.

Furthermore, Spitale et al. (2023) examined the extent to which LLMs such as GPT-3 can inform and misinform the public on important health, science, and environmental issues. In an experiment where real Twitter users were exposed to accurate information and disinformation in tweets generated by GPT-3, it was revealed that GPT-3 is a double-edged sword: it produces accurate information that is easier to understand than human-written content but also generates disinformation that is more compelling and harder to detect. Participants were better at recognizing disinformation in human-written tweets than in AI-generated ones,

while they found AI-generated accurate tweets more credible than human-written accurate tweets. Notably, humans struggled to differentiate between AI-generated and human-generated tweets, often performing no better than random guessing. These results indicate both the risks and potential of AI-generated text, emphasizing that AI-generated misinformation may be even more harmful than misinformation created by humans.

4.3.5 Comparative performance of LLMs across misinformation tasks

Supplementary Table S2 provides a comparative synthesis of LLMs reviewed in this study, summarizing their relative performance across misinformation-related tasks, including detection, generation, and mitigation. ChatGPT-4.0 emerged as a consistent top performer in both detection (Deiana et al. 2023; Wang et al. 2023) and resistance to hallucinations during content generation (McIntosh et al. 2023). However, Google Bard outperformed ChatGPT-4.0 in specific evaluation contexts. Google Bard demonstrated superior accuracy and public health alignment in fact-checking scenarios with more accessible and practical responses (Garbarino and Bragazzi 2024). Bard also demonstrated stronger mitigation performance in politically sensitive contexts, including a greater likelihood of including disclaimers or debunking (Makhortykh et al. 2024). Notably, Poe (powered by Claude 2) was the only model to consistently refuse disinformation generation across all timepoints and topics, including jailbreak attempts (Menz et al. 2024), indicating the strongest safeguards. Several studies emphasized the influence of prompt type and context on LLM performance. For example, Deiana et al. (2023) found that contextual prompting significantly improved misinformation detection accuracy. Kumar et al. (2024) demonstrated that model accuracy depended on the type of prompt (zero-shot vs. few-shot), with zero-shot tasks generally yielding higher performance. The exclusion of sentiment and emotion features further improved detection accuracy.

5 Discussion

This review examined the empirical landscape on the role of generative AI in the generation, detection, mitigation, and impact of misinformation. Across 24 studies published between 2023 and 2024, the findings present a complex and at times contradictory portrait of LLMs. These tools simultaneously pose risks as powerful misinformation generators and offer promise as scalable instruments for detection and correction. This duality underscores the urgent need for clearer guardrails, more consistent performance standards, and interdisciplinary collaboration to shape their responsible deployment.

5.1 Key findings

This review identifies four core insights regarding generative AI's role in the misinformation ecosystem. First, LLMs can generate highly credible misinformation, particularly when personalized or tailored to user identities (Gabriel et al. 2024; Kim et al. 2024; Menz et al. 2024). Second, LLM-based detection is consistent: accuracy varies by prompt phrasing, domain, and language, with weaker performance in culturally contested or non-English contexts (Deiana et al. 2023; Kumar et al. 2024; Węcel et al. 2023). Third, while mitigation strategies show promise but remain fragile. Personalized corrections can improve discernment, yet misalignment with user identity or ideology may backfire (Gabriel et al. 2024), while model safeguards are neither durable nor uniformly effective (Menz et al. 2024; Węcel et al. 2023). Finally, AI-generated misinformation influences attitudes and decisions (Kim et al. 2023; Spitale et al. 2023). Together, these findings reveal generative AI's dual potential to both exacerbate and mitigate misinformation, and underscore urgent needs for robust evaluation, contextual sensitivity, and regulatory oversight.

5.2 Theoretical implications and future directions

This review advances theoretical understanding of generative AI's role in the misinformation ecosystem by foregrounding the concept of epistemic ambivalence—the potential of AI to simultaneously construct and erode public knowledge. Such duality calls for a more critical lens on the epistemic authority of AI systems, particularly as they become central intermediaries in domains such as health, science, and politics. Drawing from sociotechnical systems theory (Kudina and van de Poel 2024), AI must be examined not merely as a technical artifact, but as a socially embedded system entangled with broader sociocultural, institutional norms, and communicative ecologies. Misinformation in LLMs, in this perspective, emerges from this entanglement, the interplay between technological design, socio-political context, and user interpretation. The "machine heuristic" further illustrates the epistemic risks, as users often overtrust AI outputs, especially when linguistically fluent or identity-congruent, thereby reinforcing motivated reasoning and reducing the likelihood of critical scrutiny. These findings resonate with prior work on misinformation vulnerability (e.g., Lewandowsky 2023; Pérez-Escolar et al. 2023), extending into the context of generative AI. This review helps position AI misinformation scholarship within a more interdisciplinary framework, calling for further integration of psychological and sociotechnical perspectives.

Building on these implications, future research should prioritize methodological rigor, contextual relevance, and global applicability. First, standardized benchmarks and protocols are needed for evaluating generative AI's performance in misinformation. The absence of shared datasets, task definitions, and reporting standards hampers comparative assessments across studies. Just as computer vision research relies on common image datasets, AI misinformation studies would benefit from shared corpora, evaluation metrics, and reporting guidelines. Second, researchers should explore how generative AI interacts with individual differences, including need for cognition, digital literacy, machine heuristics, political ideology, or cultural orientations. Understanding these dynamics is crucial for designing tailored and ethically responsible interventions. Third, research must expand beyond Western contexts. Current evidence remains Western-centric, with limited representation from Southern Asia, Africa, South America, and non-English contexts. Cultural context plays a crucial role in how people perceive, believe, and spread misinformation, as well as perceptions of AI, but it remains unclear whether LLMs perform similarly in non-Western settings or when trained and evaluated in different languages and sociopolitical environments. Fourth, longitudinal and real-world studies are needed. How does repeated exposure to AI-generated content shape attitudes, knowledge, and behaviors over time? What safeguards remain effective as users become more familiar with or skeptical of AI-generated messages? And how do these dynamics vary across platforms, cultures, and media environments? Importantly, studies should assess diverse domains—including health, elections, science, and finance—to better understand how generative AI performs in high-stakes misinformation contexts. Finally, research must attend to the structural drivers shaping AI development and deployment—including training data, corporate incentives, regulatory oversight, and democratic accountability. Otherwise, interventions may treat symptoms while leaving root causes unexamined.

5.3 Practical recommendations

We outline key practical recommendations for researchers, developers, and policymakers. First, standardized detection protocols must be developed to address inconsistencies across prompts, topics, and platforms. These protocols should be evaluated with adversarial prompt testing and include benchmarks for prompt sensitivity, using multilingual, multicultural, and domain-diverse datasets that reflect the global information environment, beyond English-language and Western-centric content. Second, safety guardrails require continuous validation and longitudinal monitoring. Evaluations should be documented publicly, with transparent reporting of failures and updates in safety performance. Third, user-facing safeguards should be considered to promote critical engagement with AI outputs, including interface-level cues, such as warning labels or fact-check toggles. Fourth, scalable content attribution tools, such as digital watermarking, embedded metadata, or AI-generated content labels, should be adopted across platforms to enhance transparency and mitigate the spread of AI-generated misinformation. Lastly, policymakers and institutional actors must establish and enforce shared standards for content provenance, safety audits, and jailbreak prevention. Regulatory oversight, coupled with cross-sector collaboration among industry, academia, and civil society, is essential to ensure that generative AI systems align with democratic values and serve the public interest.

5.4 Limitations

Several limitations should be acknowledged. First, our inclusion criteria focused on publicly available generative AI systems, excluding highly customized or proprietary models. Future reviews might expand the scope to include both commercial and non-commercial models to capture a more complete picture of the technological landscape. Second, although the review identifies broad themes and trends, it does not quantify effect sizes or aggregate findings statistically. Given the diversity of research designs and outcome measures, a scoping review was appropriate for mapping the field, but it limits the ability to make comparative or evaluative judgments about intervention effectiveness or model performance. Third, the scope of the review was limited to studies published in English and retrieved primarily through Google Scholar. This search may underrepresent certain fields, non-English languages, or less prominent venues. Most reviewed studies were conducted in North America, Europe, and Australia, with limited representation from Asia, Africa, or South America. This geographic concentration introduces a Western-centric bias, raising concerns about the global generalizability of the findings. Finally, claims about the impact of AI-generated misinformation are largely based on a narrow set of experimental studies, many centered on travel-related contexts (e.g., Kim et al. 2023). These insights are valuable but may not generalize to broader domains.

6 Conclusion

This review underscores the paradoxical role of generative AI in the misinformation ecosystem. On the one hand, LLMs have demonstrated a remarkable capacity to generate and personalize misinformation at scale, raising serious concerns for public trust and democratic resilience. On the other hand, they offer new tools for detection, correction,

and prevention—tools that, if properly developed and governed, could serve as a bulwark against the very threats they help create. Navigating this tension will require not only continued empirical research but also thoughtful, multidisciplinary engagement with the broader sociotechnical systems in which generative AI operates.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00146-025-02620-3.

Author contributions S.P. and X.N. jointly conceived the study and designed the review protocol. S.P. conducted the initial literature search and data extraction. Both authors collaboratively analyzed and interpreted the data. All authors wrote the main manuscript text and S.P. prepared tables and figures. X.N. reviewed and approved the final manuscript.

Data availability The data underlying this article, including the list of publications included in the review, are available within the article.

Declarations

Conflict of interest The authors declare no competing interests.

References

*Publications included in the review

Athaluri SA, Manthena SV, Kesapragada VKM, Yarlagadda V, Dave T, Duddumpudi RTS (2023) Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus 15:Article e37432. https://doi.org/10.7759/cureus.37432

Augenstein I, Baldwin T, Cha M et al (2024) Factuality challenges in the era of large language models and opportunities for fact-checking. Nat Mach Intell 6:852–863. https://doi.org/10.1038/s42256-024-00881-z

Aydın Ö, Karaarslan E (2023) Is ChatGPT leading generative AI? What is beyond expectations? Acad Platf J Eng Smart Syst 11:118–134. https://doi.org/10.21541/apjess.1293702

Bandara C (2024) Hallucination as disinformation: the role of LLMs in amplifying conspiracy theories and fake news. J Appl Cybersecur Anal Intell Decis-Mak Syst 14:65–76

Cascella M, Montomoli J, Bellini V, Bignami E (2023) Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 47:Article 33. https://doi.org/10.1007/s10916-023-01925-4

Chan MPS, Jones CR, Hall Jamieson K, Albarracín D (2017) Debunking: a meta-analysis of the psychological efficacy of messages countering misinformation. Psychol Sci 28:1531–1546. https://doi.org/10.1177/0956797617714579

Costello TH, Pennycook G, Rand DG (2024) Durably reducing conspiracy beliefs through dialogues with AI. Science 385:Article eadq1814. https://doi.org/10.1126/science.adq1814

*Deiana G, Dettori M, Arghittu A, Azara A, Gabutti G, Castiglia P (2023) Artificial intelligence and public health: evaluating ChatGPT responses to vaccination myths and misconceptions. Vaccines 11:Article 1217. https://doi.org/10.3390/vaccines11071217

Del Vicario M, Bessi A, Zollo F et al (2016) The spreading of misinformation online. Proc Natl Acad Sci USA 113:554–559. https://doi.org/10.1073/pnas.1517441113

Di Domenico G, Ding Y (2023) Between brand attacks and broader narratives: How direct and indirect misinformation erode consumer trust. Curr Opin Psychol 54:Article 101716. https://doi.org/10.1016/j.copsyc.2023.101716

Dobber T, Metoui N, Trilling D, Helberger N, De Vreese C (2021) Do (microtargeted) deepfakes have real effects on political attitudes? Int J Press Polit 26:69–91. https://doi.org/10.1177/1940161220944364

Ecker UKH, Lewandowsky S, Cook J, Schmid P, Fazio LK, Brashier N, Kendeou P, Vraga EK, Amazeen MA (2022) The psychological drivers of misinformation belief and its resistance to correction. Nat Rev Psychol 1:13–29. https://doi.org/10.1038/s44159-021-00006-y

Ecker UK, Tay LQ, Roozenbeek J, Van Der Linden S, Cook J, Oreskes N, Lewandowsky S (2024) Why misinformation must not be ignored. Am Psychol. https://doi.org/10.1037/amp0001448

*Gabriel S, Lyu L, Siderius J, Ghassemi M, Andreas J, Ozdaglar A (2024) Generative AI in the era of 'alternative facts.' An MIT exploration of generative AI. https://doi.org/10.21428/e4baedd9.82175d26

*Garbarino S, Bragazzi NL (2024) Evaluating the effectiveness of artificial intelligence-based tools in detecting and understanding sleep health misinformation: comparative analysis using Google Bard and OpenAI ChatGPT-4. J Sleep Res 33:Article e14210. https://doi.org/10.1111/jsr.14210

Huang J, Chang KCC (2022) Towards reasoning in large language models: a survey. PsyArXiv. https://doi.org/10.48550/arXiv.2212.10403

Ireton C, Posetti J (2018) Journalism, fake news & disinformation: handbook for journalism education and training. UNESCO, Paris

Kamel H (2024) Understanding the impact of AI Hallucinations on the university community. Cybrarians J 73:111–134. https://doi.org/10.70000/cj.2024.73.622

*Kim JH, Kim J, Park J, Kim C, Jhang J, King B (2023) When ChatGPT gives incorrect answers: the impact of inaccurate information by generative AI on tourism decision-making. J Travel Res 64:51–73. https://doi.org/10.1177/00472875231212996

*Kim JJ, Srivatsa AV, Nahas GR et al (2024) Generative AI can effectively manipulate data. AI Ethics. https://doi.org/10.1007/s43681-024-00546-y

Kudina O, van de Poel I (2024) A sociotechnical system perspective on AI. Minds Mach 34:Article 21. https://doi.org/10.1007/s11023-024-09680-2

*Kumar R, Goddu B, Saha S, Jatowt A (2024) Silver lining in the fake news cloud: can large language models help detect misinformation?. IEEE Trans Artif Intell 6:14–24. https://doi.org/10.1109/TAI.2024.3440248

Kuznetsova E, Makhortykh M, Vziatysheva V, Stolze M, Baghumyan A, Urman A (2025) In generative AI we trust: can chatbots

effectively verify political information? J Comput Soc Sci 8:Article 15. https://doi.org/10.1007/s42001-024-00338-8
Lewandowsky S (2023) Demagoguery, technology, and cognition: Addressing the threats to democracy. In: Hatzivassiliou E (ed) Digital technologies and the stakes for representative democracy. Alpha Omega Publishing, Athens, pp 83–93. https://www.parlamento.pt/Documents/2023/junho/Digital-Technologies-Stakes-Representative-Democracy-Athens-June-2022.pdf
Lu H (2025) Generative AI for vaccine misbelief correction: insights from targeting extraversion and pseudoscientific beliefs. Vaccine 54:Article 127018. https://doi.org/10.1016/j.vaccine.2025.127018
*Makhortykh M, Sydorova M, Baghumyan A, Vziatysheva V, Kuznetsova E (2024) Stochastic lies: How LLM-powered chatbots deal with Russian disinformation about the war in Ukraine. Harvard Kennedy School (HKS) Misinf Rev. https://doi.org/10.37016/mr-2020-154
*McIntosh TR, Liu T, Susnjak T, Watters P, Ng A, Halgamuge MN (2023) A culturally sensitive test to evaluate nuanced GPT hallucination. IEEE Trans Artif Intell 5:2739–2751. https://doi.org/10.1109/TAI.2023.3332837
*McMahon HV, McMahon BD (2024) Automating untruths: ChatGPT, self-managed medication abortion, and the threat of misinformation in a post-Roe world. Front Digit Health 6:Article 1287186. https://doi.org/10.3389/fdgth.2024.1287186
*Menz BD, Kuderer NM, Bacchi S et al (2024) Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. BMJ 384:Article e078538. https://doi.org/10.1136/bmj-2023-078538
Monteith S, Glenn T, Geddes JR, Whybrow PC, Achtyes E, Bauer M (2024) Artificial intelligence and increasing misinformation. Br J Psychiatry 224:33–35. https://doi.org/10.1192/bjp.2023.136
Muhammed TS, Mathew SK (2022) The disaster of misinformation: a review of research in social media. Int J Data Sci Anal 13:271–285. https://doi.org/10.1007/s41060-022-00311-6
Nan X, Iles IA, Yang B, Ma Z (2022a) Public health messaging during the COVID-19 pandemic and beyond: lessons from communication science. Health Commun 37:1–19. https://doi.org/10.1080/10410236.2021.1994910
Nan X, Wang Y, Thier K (2022b) Why do people believe health misinformation and who is at risk? A systematic review of individual differences in susceptibility to health misinformation. Soc Sci Med 314:Article 115398. https://doi.org/10.1016/j.socscimed.2022.115398
Nan X, Thier K, Wang Y (2023) Health misinformation: what it is, why people believe it, how to counter it. Ann Int Commun Assoc 47:381–410. https://doi.org/10.1080/23808985.2023.2225489
Neely SR, Eldredge C, Ersing R, Remington C (2022) Vaccine hesitancy and exposure to misinformation: a survey analysis. J Gen Intern Med 37:179–187. https://doi.org/10.1007/s11606-021-07171-z
Noar SM, Francis DB, Bridges C, Sontag JM, Ribisl KM, Brewer NT (2016) The impact of strengthening cigarette pack warnings: systematic review of longitudinal observational studies. Soc Sci Med 164:118–129. https://doi.org/10.1016/j.socscimed.2016.06.011
Ognyanova K, Lazer D, Robertson RE, Wilson C (2020) Misinformation in action: fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. Harvard Kennedy School (HKS) Misinf Rev. https://doi.org/10.37016/mr-2020-024
Pérez-Escolar M, Lilleker D, Tapia-Frade A (2023) A systematic literature review of the phenomenon of disinformation and misinformation. Media Commun 11:76–87. https://doi.org/10.17645/mac.v11i2.6453
Romero Moreno F (2024) Generative AI and deepfakes: a human rights approach to tackling harmful content. Int Rev Law Comput Technol 38:297–326. https://doi.org/10.1080/13600869.2024.2324540
*Santangeli A, Mammola S, Nanni V, Lambertucci SA (2024) Large language models debunk fake and sensational wildlife news. Integr Conserv 3:127–133. https://doi.org/10.1002/inc3.55
Schmid P, Altay S, Scherer LD (2023) The psychological impacts and message features of health misinformation. Eur Psychol 28:162–172. https://doi.org/10.1027/1016-9040/a000494
*Senekal B, Brokensha S (2023) Is ChatGPT a friend or foe in the war on misinformation? A South African perspective. Communicare J Commun Sci South Africa 42:3–16. https://hdl.handle.net/10520/ejc-comcare_v42_n2_a3
*Sparks CA, Fasulo SM, Windsor JT, Bankauskas V, Contrada EV, Kraeutler MJ, Scillia AJ (2024) ChatGPT is moderately accurate in providing a general overview of orthopaedic conditions. JBJS Open Access 9:Article e23. https://doi.org/10.2106/JBJS.OA.23.00129
*Spitale G, Biller-Andorno N, Germani F (2023) AI model GPT-3 (dis) informs us better than humans. Sci Adv 9:Article eadh1850. https://doi.org/10.1126/sciadv.adh1850
Suarez-Lledo V, Alvarez-Galvez J (2021) Prevalence of health misinformation on social media: systematic review. J Med Internet Res 23:Article e17187. https://doi.org/10.2196/17187
Sun Y, Sheng D, Zhou Z, Wu Y (2024) AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanit Soc Sci Commun 11:1–14. https://doi.org/10.1057/s41599-024-03811-x
Sundar SS (2008) The MAIN model: A heuristic approach to understanding technology effects on credibility. In: Metzger MJ, Flanagin AJ (eds) Digital media, youth, and credibility. The MIT Press, Cambridge, pp 73–100. https://doi.org/10.1162/dmal.9780262562324.073
Swire-Thompson B, Lazer D (2020) Public health and online misinformation: challenges and recommendations. Annu Rev Public Health 41:433–451. https://doi.org/10.1146/annurev-publhealth-040119-094127
Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J (2020) Deepfakes and beyond: a survey of face manipulation and fake detection. Inf Fusion 64:131–148. https://doi.org/10.1016/j.inffus.2020.06.014
Traberg CS, Roozenbeek J, van der Linden S (2022) Psychological inoculation against misinformation: current evidence and future directions. Ann Am Acad Polit Soc Sci 700:136–151. https://doi.org/10.1177/00027162221087936
Tricco AC, Lillie E, Zarin W et al (2018) PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 169:467–473. https://doi.org/10.7326/M18-0850
Vaccari C, Chadwick A (2020) Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Soc Media + Soc 6:Article 056305120903408. https://doi.org/10.1177/2056305120903408
Van der Linden S, Leiserowitz A, Rosenthal S, Maibach E (2017) Inoculating the public against misinformation about climate change. Glob Chall 1:Article 1600008. https://doi.org/10.1002/gch2.201600008
Verdoliva L (2020) Media forensics and deepfakes: an overview. IEEE J Sel Top Signal Process 14:910–932. https://doi.org/10.1109/JSTSP.2020.3002101
Verma N (2024) "One Video Could Start a War": a qualitative interview study of public perceptions of Deepfake Technology. Proc Assoc Inf Sci Technol 61:374–385. https://doi.org/10.1002/pra2.1035
Vraga EK, Bode L (2020) Defining misinformation and understanding its bounded nature: using expertise and evidence for describing

misinformation. Polit Commun 37:136–144. https://doi.org/10.1080/10584609.2020.1716500

Walter N, Tukachinsky R (2020) A meta-analytic examination of the continued influence of misinformation in the face of correction: how powerful is it, why does it happen, and how to stop it? Commun Res 47:155–177. https://doi.org/10.1177/0093650219854600

Wang Y, McKee M, Torbica A, Stuckler D (2019) Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 240:Article 112552. https://doi.org/10.1016/j.socscimed.2019.112552

*Wang G, Gao K, Liu Q, Wu Y, Zhang K, Zhou W, Guo C (2023) Potential and limitations of ChatGPT 3.5 and 4.0 as a source of COVID-19 information: comprehensive comparative analysis of generative and authoritative information. J Med Internet Res 25:Article e49771. https://doi.org/10.2196/49771

*Węcel K, Sawiński M, Stróżyna M, Lewoniewski W, Księżniak E, Stolarski P, Abramowicz W (2023) Artificial intelligence-friend or foe in fake news campaigns. Econ Bus Rev 9:41–70. https://doi.org/10.18559/ebr.2023.2.736

Westerlund M (2019) The emergence of Deepfake Technology: a review. Technol Innov Manag Rev 9:40–53. https://doi.org/10.22215/timreview/1282

Zhou T, Li S (2024) Understanding user switch of information seeking: from search engines to generative AI. J Librariansh Inf Sci. https://doi.org/10.1177/09610006241244800

Zimmerman T, Shiroma K, Fleischmann KR, Xie B, Jia C, Verma N, Lee MK (2023) Misinformation and COVID-19 vaccine hesitancy. Vaccine 41:136–144. https://doi.org/10.1016/j.vaccine.2022.11.014