Article

Not peer-reviewed version

Large Language Models and Social Media Information Integrity: Opportunities, Challenges, and Research Directions

Junjie Xiong *, Zhengyuan Jiang, Xiaoran Xu, Chi Zhang, Changjia Zhu, Ning Wang, Mingkui Wei, Zhuo Lu, Yao Liu, Lingyao Li *

Posted Date: 19 August 2025

doi: 10.20944/preprints202508.1280.v1

Keywords: large language models; social media; security implications; misinformation; disinformation; fake news; LLM-enhanced bots; privacy

 doi:10.20944/preprints202508.1280.v1

Article

Large Language Models and Social Media Information Integrity: Opportunities, Challenges, and Research Directions

Junjie Xiong [1,*], Zhengyuan Jiang [2], Xiaoran Xu [2], Chi Zhang [2], Changjia Zhu [2], Ning Wang [2], Mingkui Wei [3], Zhuo Lu [2], Yao Liu [2] and Lingyao Li [2,*]

[1] Missouri University of Science and Technology, Rolla, U.S.
[2] University of South Florida, Tampa, U.S.
[3] George Mason University, Fairfax, U.S.
* Correspondence: junjiexiong@mst.edu (J.X.); lingyaol@usf.edu (L.L.)

Abstract: Large Language Models (LLMs) have emerged as powerful tools that impact information integrity on social media platforms. This comprehensive review examines the dual role of LLMs in both facilitating and mitigating various information integrity challenges, including misinformation, disinformation, fake news, social bots, and privacy concerns. Through a systematic analysis of papers from academic databases, we identify key patterns in how LLMs influence social media ecosystems. Our findings reveal that while LLMs can enhance detection capabilities for malicious content and enable sophisticated defense mechanisms, they simultaneously pose risks by enabling the generation of highly convincing, deceptive content. We categorize and analyze the potential and challenges across different dimensions of information integrity, examining technical capabilities, ethical implications, and privacy concerns. The study demonstrates critical gaps in current approaches, particularly in cross-lingual detection, real-time monitoring, and privacy-preserving implementations. We conclude by proposing future research directions and recommendations for stakeholders to leverage LLMs while mitigating risks in social media information integrity.

Keywords: large language models; social media; security implications; misinformation; disinformation; fake news; LLM-enhanced bots; privacy

1. Introduction

Social media platforms have become indispensable tools for modern communication, connecting users around the world and facilitating real-time information sharing. However, these digital platforms remain vulnerable to critical information security threats, including misinformation, disinformation, and the propagation of biased or toxic content [57,152]. These vulnerabilities not only erode public trust but also present substantial challenges in maintaining the integrity of digital systems. The growing risks associated with the dissemination of false information and cyber threats underscore the urgent need to safeguard online environments [154].

Large Language Models (LLMs) have rapidly become embedded in today's digital platforms. These AI systems—exemplified by GPT-style models—drive a range of applications from conversational agents to intelligent digital assistants and content creation tools. Social media platforms increasingly leverage LLM-driven bots to engage users or automate customer service, while writing assistants use them to generate articles, marketing copy, and code. The integration of LLMs is reshaping how information is produced and consumed in digital systems. On the positive side, LLMs have shown potential in detecting and mitigating harmful content [86,167]. LLM-powered applications can also help moderate content and flag policy violations at scale, complementing human moderators in keeping online communities civil [48,63]. These applications show that LLMs carry significant implications for the quality of public discourse and the security of digital environments.

 doi:10.20944/preprints202508.1280.v1

Despite their capabilities, the integration of LLMs can also impact information integrity. A major concern is their tendency to "hallucinate," generating text that appears authentic and authoritative but may be false or misleading [31,139]. For example, recent studies show that LLMs can produce election-related misinformation that is almost indistinguishable from human-written content [18,145]. Additionally, they have been known to generate misinformation or fake news, including hateful or offensive language, particularly when trained on toxic datasets or manipulated through jailbreak attacks [63,93]. High-profile incidents, such as an AI chatbot coerced into spreading racist and offensive messages, illustrate this risk. Moreover, because LLMs learn from vast internet datasets, they can internalize and reinforce societal biases, associating certain groups with negative stereotypes and potentially harming marginalized communities [44,64,93]. When combined with social media's amplification mechanisms, LLM-generated misinformation or harmful content can spread rapidly, misleading users on critical topics and further complicating efforts to maintain the integrity of digital discourse.

To mitigate these issues, researchers and practitioners have proposed a variety of mitigation strategies to align LLM behavior with ethical and safety standards [64,68]. One typical approach is the reinforcement learning from human feedback (RLHF) alignment pipeline, which fine-tunes models using human input to reinforce desirable behaviors. By training on examples of preferred behaviors and using human judgments to reward desirable answers, LLMs can be steered toward more helpful and harmless responses. In addition, researchers have developed data curation pipelines to filter out misinformation from training corpora [146]. Such dataset-level interventions can preempt many issues before the model engages with users. Another typical strategy is the adversarial testing [48]. For example, adversarial prompts are designed to find weaknesses. By learning from these adversarial cases (e.g., either via additional fine-tuning or adjusting the model's safety filters), developers can strengthen the model against malicious exploitation. These typical strategies can help ensure that known failure modes are addressed before online users encounter them.

While prior studies have shown that LLMs can both contribute to and help mitigate online information security risks, there remains a lack of comprehensive review papers that synthesize these findings. In addition, most existing research focuses on specific aspects of these issues, such as adversarial attacks, bias mitigation, or ethical AI, without offering a perspective on the broader implications within the context of social media platforms. Given the rapid advancements in LLMs and the increasing integration of LLMs into social media platforms, a systematic review is necessary to summarize current knowledge, identify research gaps, and provide a structured analysis of potential risks and mitigation strategies. Therefore, this study seeks to address the following research questions:

- **RQ1.** What is the potential of LLMs in enhancing social media information integrity?
- **RQ2.** How do LLMs challenge the detection and mitigation of information integrity issues?
- **RQ3.** To what extent do ethical and security implications emerge from LLMs' role in information integrity?

This paper makes several key contributions to the understanding of LLMs in social media information integrity. First, we provide a comprehensive framework for analyzing information security issues in social media platforms and explain how LLMs have become involved in these challenges. Second, through a systematic review of recent literature, we identify and categorize the key potentials of LLMs in enhancing information integrity, supported by empirical evidence and practical implementations. Third, we present a detailed analysis of the challenges and limitations that need to be addressed when deploying LLMs in social media environments. Finally, we propose future research directions and practical recommendations for stakeholders to leverage LLMs effectively while mitigating associated risks.

 doi:10.20944/preprints202508.1280.v1

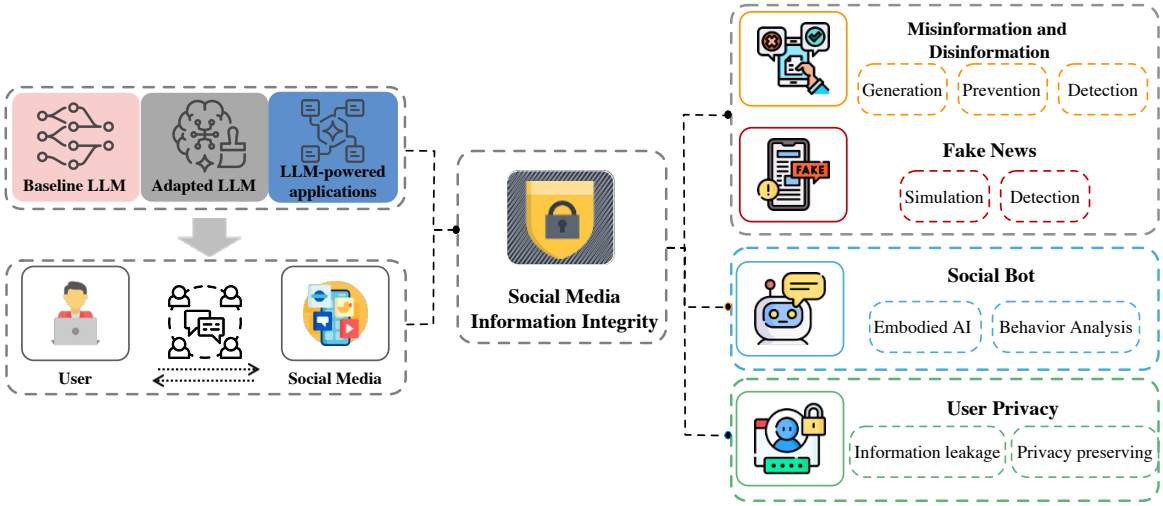

Figure 1. Framework of this survey: The Integration and Impact of LLMs in Social Media Information Integrity. At the center, LLMs, as powerful tools for social media content creation, can both strengthen and compromise information integrity depending on their application. On the one side, malicious actors can exploit LLMs to generate sophisticated misinformation and disinformation, create convincing fake news, deploy advanced social bots, and compromise user privacy through automated data extraction and profiling. On the other side, LLMs can be employed to detect and counter these threats through content verification, bot detection, privacy protection, and automated fact-checking systems. The bidirectional arrows indicate the dynamic nature of the technological arms race, where advances in both attack and defense continuously evolve.

2. Backgrounds

2.1. What Are the Information Security Issues in Social Media?

The challenge of false information has become increasingly critical as social media platforms emerge as primary news sources, with reports indicating that 54% of US adults now obtain their news through platforms like Facebook and YouTube [24]. Our analysis of social media security issues reveals three interconnected categories that represent the primary challenges in maintaining information integrity. First, **information disorder**—encompassing misinformation, disinformation, and fake news—represents the content-level challenges. This category includes both unintentional spread of incorrect information and deliberate manipulation of facts, with studies showing that pseudoscience, conspiracy theories, lies, and deepfakes constitute significant threats [7]. Second, **social bots** represent the agent-level challenges, serving as primary vectors for amplifying and automating the spread of information integrity issues. These automated systems, enhanced by LLMs, can now generate highly convincing content while mimicking human behavior patterns [146]. Third, **user privacy** concerns represent the infrastructure-level challenges, where the need to protect personal information intersects with the requirements for effective content moderation and bot detection [114]. The advent of LLMs has significantly amplified these challenges across all three categories. For example, models like GPT-3 can generate highly convincing false information that humans find both comprehensible and compelling [145]. This capability becomes particularly concerning in critical domains such as health-related topics [19,140], where misinformation about allergies, abortion suggestions, and other medical procedures can have serious consequences, even in the presence of platform safeguards such as TikTok's misinformation.

 doi:10.20944/preprints202508.1280.v1

Table 1. Definitions and Sources of Key Information Integrity Concepts in Social Media.

Category	Author and Year	Definition	Origin
Misinformation	Altay et al. (2023)[7]	False and misleading information.	Academia
	Wardle et al.(2017)[159]	Information that is false but not created with the intention of causing harm.	Academia
	UNHCR (2021)[155]	Misinformation is false or inaccurate information. Examples include rumors, insults, and pranks.	NGO
Disinformation	Wardle et al.(2017)[159]	Information that is false and is knowingly shared to cause harm.	Academia
	UNHCR (2021)[155]	Disinformation is deliberate and includes malicious content such as hoaxes, spear phishing, and propaganda. It spreads fear and suspicion among the population.	NGO
Fake News	Cooke et al.(2017) [36]	False and often sensational information disseminated under the guise of news reporting, yet the term has evolved over time and has become synonymous with the spread of false information	Academic
	Allcott et al. (2017) [5]	News articles that are intentionally and verifiably false and could mislead readers.	Academic
Social Bot	Staab et al.(2023) [146] Lyu et al.(2023) [100]	Social bots capable of inferring and utilizing multimodal capabilities, effectively processing and generating both textual and visual content.	Academic
	Cloudflare (2025) [32]	Social bots are automated programs that mimic human users, operating partially or fully on their own, with many being used for deceptive or harmful purposes.	Industry
User Privacy	Helen et al.(2019) [114]	The right of users to control information flows and disclosures about themselves according to contextual norms and social expectations.	Academia
	Facebook (2023) [105]	Providing users transparency and practical control over how their personal data is collected, used, and shared, consistent with regulatory compliance.	Industry

Information Disorder, defined in Shu et al. [142], Wardle and Derakhshan [159], includes false information intended to harm and factual information being used to manipulate. In the context of social media, we focus on the harmful aspects, known as **misinformation**, **disinformation**, and **fake news**. LLM can further complicate the information ecosystem. On one hand, LLMs can be exploited to automatically generate misleading content, impersonate users, or flood platforms with synthetic narratives, thereby intensifying the spread of information disorder. On the other hand, these models also offer potential countermeasures: they can be fine-tuned to detect falsehoods, identify manipulation patterns, and assist in fact-checking at scale. Misinformation and Disinformation represent two distinct but related threats to information integrity. Misinformation refers to false information shared without malicious intent [159], while disinformation involves deliberately created false content intended to cause harm [159]. The United Nations High Commissioner for Refugees (UNHCR) further distinguishes these two concepts by emphasizing that disinformation includes specific malicious content types such as hoaxes and propaganda [155].

Fake News has overlaps with misinformation and disinformation, but specifically focuses on false information presented in a news format. Fake news is distinguished from other forms of false information through three key characteristics [5,36]: its deliberate mimicry of legitimate news media formatting and style; its exploitation of established news distribution channels; and its intent to deceive by leveraging the credibility traditionally associated with journalism. Fake news refers to objectively false information presented in various forms, including news articles, public statements, speeches, and social media posts [182]. Its core characteristic lies in the deviation from objective facts, regardless of the publisher's motivation [156]. Whether created for malicious deception, entertainment, satire,

 doi:10.20944/preprints202508.1280.v1

or through unintentional dissemination, any content that contradicts factual truth falls under this category. This phenomenon has a long history of influencing political views and social discourse [85].

Social Bots are automated programs designed to mimic human interaction on social media platforms, which have evolved significantly with the advent of LLMs. Recent research underscores their enhanced multimodal capabilities [100,146], while industry perspectives from Cloudflare emphasize their potential for deceptive purposes [32]. Traditional social bots—including early chatbots, spiders, and coordinated fake accounts—have long shaped the information landscape through automated engagement and manipulation. These earlier systems operated on predefined scripts and rules, lacking contextual adaptation [53,60].

Social bots enhanced by LLM represent a significant evolution in automated social media manipulation. As documented by Li et al. [87], these advanced systems leverage LLMs with billions of parameters, demonstrating unprecedented abilities in natural language understanding and generation. These bots exhibit sophisticated camouflage abilities and can generate highly contextual responses that closely mimic human behavior [167]. The integration of LLMs has fundamentally transformed bot capabilities in several key aspects. Feng et al. [48] demonstrated that these bots can now generate highly personalized content while adapting their behavior through continuous learning from user interactions. This advancement is particularly significant in social engineering attacks, where LLM-enhanced bots can craft convincing phishing content that exploits contextual understanding [34]. Furthermore, Staab et al. [146] revealed that these bots can infer and utilize personal attributes from text with up to 85% accuracy, making their interactions increasingly sophisticated and targeted. Recent studies have also pointed out the multimodal capabilities of LLM-enhanced bots. Lyu et al. [100] showed that these bots can effectively process and generate both textual and visual content while maintaining coherent context across different modalities. This evolution in bot capabilities not only introduces new challenges for detection systems due to their increasingly sophisticated and human-like interactions, but also opens promising opportunities for enhancing information integrity through more adaptive, context-aware bot detection and moderation frameworks [54].

User Privacy encompasses users' rights to control their information flows within specific contexts [114]. This definition aligns with but extends beyond industry standards, including Facebook's emphasis on transparency and regulatory compliance [105]. Integrating LLMs into social media platforms has introduced unprecedented privacy challenges that demand systematic analysis and innovative solutions. These challenges primarily manifest in data exposure and inference capabilities: Staab et al. [146,146] reveal how LLMs can breach privacy by deducing sensitive personal information from pseudonymized content and reconstructing detailed user profiles from seemingly unrelated pieces of information. This fundamental vulnerability extends to data memorization and leakage, where Kandpal et al. [74] and Duan et al. [41] demonstrate how LLMs can retain and expose sensitive training data. The cross-platform privacy challenge is particularly pressing, as Treves et al. [153] exposes how tools like RURLMAN can breach user privacy by automatically linking identities across multiple platforms, while Ayoobi et al. [12] identifies how these models enable the creation of deceptive profiles for malicious purposes. User Privacy risks have become increasingly complex in cross-platform and real-time deployment scenarios. Previous research [20,22,29,40,50,112,120,158,174,175] has demonstrated that model inversion and membership inference attacks can effectively reconstruct private training data or infer user participation based on social media API outputs. Even though state-of-the-art techniques have shown the potential to preserve user privacy under the LLM-empowered social media environment, these vulnerabilities are particularly concerning in real-time interactions, where behavioral patterns may inadvertently reveal sensitive information such as personal habits, health status, or relationships.

To establish a foundation for our study, we present key terms and their definitions from various authoritative sources across academia, industry, and non-governmental organizations (Table 1). These definitions are crucial for understanding the multifaceted challenges that LLMs present to social media information integrity.

 doi:10.20944/preprints202508.1280.v1

2.2. Why Does LLM Raise Information Security Issues in Social Media?

The rapid evolution of social media platforms has fundamentally transformed how information is created, shared, and consumed online [57,152]. Within this landscape, LLMs have emerged as transformative technologies that significantly impact information security dynamics [146,149]. Their deep integration into social media systems, from content moderation to user interaction, has created new opportunities and challenges that demand careful examination [22,39]. This involvement stems from several fundamental factors:

Technical Capabilities and Limitations: LLMs' advanced natural language processing capabilities enable them to generate highly convincing content that can be indistinguishable from human-written text [167], while exhibiting tendencies to hallucinate [139]. This dual-use characteristic manifests in their ability to serve both as powerful detection tools, achieving accuracy rates exceeding 98% [54], and as potential threats when misused, with LLM-enhanced bots achieving up to 29.6% evasion rates [48]. Their language understanding and multilingual abilities [116] further amplify both these protective and threatening aspects, creating an ongoing technological arms race between security measures and evolving threats.

Scale and Privacy Implications: The integration of LLMs enables automated content generation at unprecedented scales [167], while raising substantial privacy concerns through their sophisticated inference capabilities. These models can extract and correlate personal information from seemingly unrelated data points, potentially revealing sensitive attributes such as location, demographics, and behavioral patterns [146]. The risk is further amplified by LLMs' ability to process and analyze vast amounts of historical user data, where patterns in content creation and interaction histories may inadvertently expose private information. This creates fundamental tensions between leveraging LLMs' powerful functionalities for content generation and ensuring robust privacy protection [121]. These characteristics make LLMs central to both the problems and solutions in social media information security. Understanding these fundamental factors is crucial for developing effective strategies to harness LLMs' potential while mitigating their risks. In the following sections, we present a comprehensive scope review of both the opportunities and challenges that emerge from LLMs' integration into social media information security systems.

In this paper, we identify three critical areas that warrant in-depth analysis in social media security: information disorder, social bots, and privacy concerns. The first area examines misinformation and disinformation detection using enhanced LLM capabilities and automated systems. The second area focuses on LLM-enhanced detection techniques for identifying automated accounts and coordinated networks. The third area addresses privacy-preserving mechanisms while ensuring information integrity. These interconnected areas represent the key challenges where LLMs present both risks and opportunities in social media environments.

3. Data, Methods and Initial Findings

3.1. Data Preparation

We select OpenAlex as our primary source for collecting relevant studies for two key reasons. First, OpenAlex is an open-access repository of scholarly metadata that indexes both preprints and peer-reviewed publications [119], allowing us to include cutting-edge research that is not only published in top-tier conferences and journals but also includes work that may not yet be formally published. This is especially beneficial in the rapidly evolving field of LLMs, where preprints are prevalent. Second, OpenAlex offers unrestricted access to its data, ensuring that our methodology remains transparent and replicable by the broader research community.

To identify relevant literature, we query OpenAlex using keywords from three core categories: "social media," "online platform," and "large language model." The full data preparation and filtering process is illustrated in Figure 2. We retrieve 1,048 papers matching our keyword criteria during this process. After removing duplicates—both identical titles and redundant formats (e.g., preprint and published versions of the same work)—we retain 956 unique records. Next, we apply a multi-stage

 doi:10.20944/preprints202508.1280.v1

screening process to ensure the remaining records align with the focus of our study. This involved three key filters: (1) **Topic Matching**: we retain papers addressing our core themes, including how LLMs facilitate information integrity, mitigate information integrity challenges, and ethical or security implications on social media field; (2) **Outdated Model Filtering**: we exclude studies not focusing on LLM, ensuring a focus on state-of-the-art LLMs; (3) **Paper Quality Verification**: we select papers that met our inclusion criteria of novelty and technical impact in addressing the capabilities or challenges of LLMs in social media contexts.

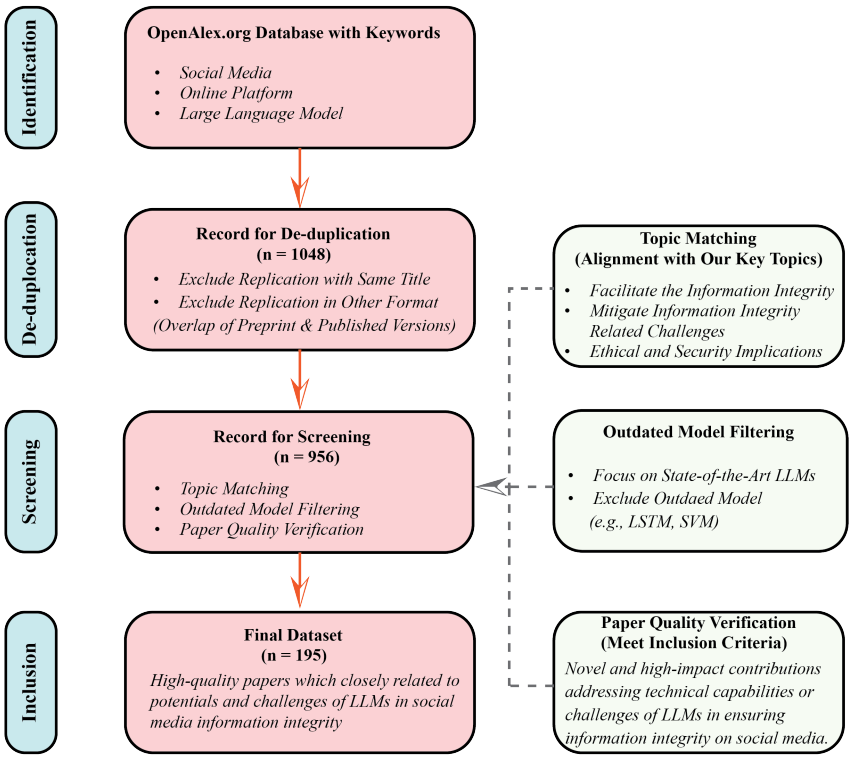

Figure 2. Data Preparation Flow.

3.2. Data Description

To comprehensively understand the current landscape of LLMs in social media integrity research, our analysis focuses on four key dimensions: model adoption (types and scales of LLMs being utilized), architectural evolution (development in model structures and capabilities), research focus areas (primary integrity challenges being addressed), and platform influence (impact across different social media platforms). These dimensions are selected to capture both the technical advancement of LLM applications and their practical impact on social media integrity. The data description of the extracted information is presented in Figure 3.

Model Adoption and Evolution: The model popularity analysis (Figure 3) shows a high concentration of research using GPT-series models in information integrity studies. This preference may be attributed to several factors: GPT models' accessibility through well-documented APIs, their extensive pre-training on diverse datasets, and their demonstrated capabilities in language understanding tasks. The growing adoption of alternative models like LLaMA and Bloom might reflect researchers' interest in open-source alternatives and specialized architectures that can be fine-tuned for specific integrity tasks. This diversification in model selection suggests an evolving research landscape where different models may serve complementary roles in addressing various aspects of information integrity.

Architectural Trends: Temporal analysis of model architectures (Figure 3) shows a consistent increase in LLM adoption across different architectural paradigms. This growth may be attributed to several factors. Decoder-only architectures like the GPT series likely gained popularity due to their

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 August 2025 doi:10.20944/preprints202508.1280.v1

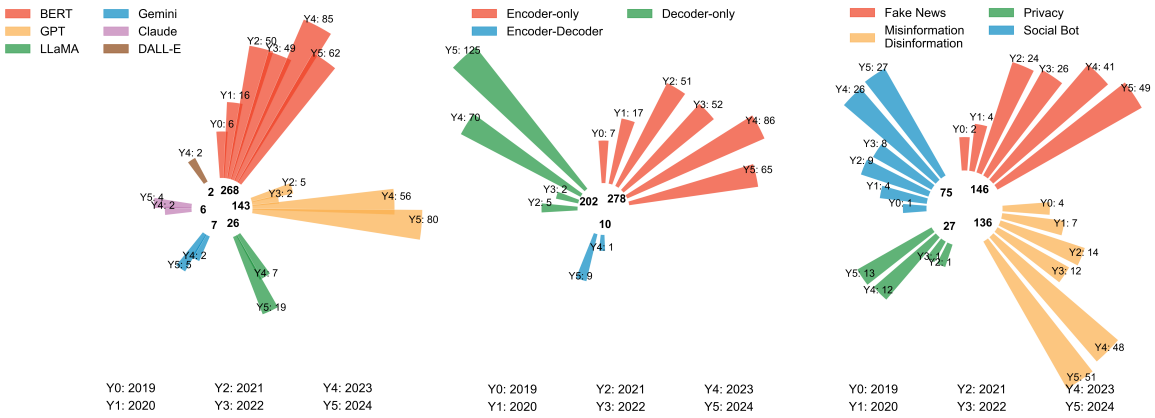

Figure 3. Data Processing Analysis of social media integrity research involving LLMs over recent years, illustrating popular models, architectural trends, and evolving research topics. 1) Model popularity: Visualization of the yearly prevalence of different LLMs (e.g., BERT, GPT, LLaMA, Claude) in social media integrity research, emphasizing temporal trends in model adoption. 2) Architectural trend: Temporal breakdown of encoder-only, decoder-only, and encoder-decoder LLM architectures, reflecting the shifting computational paradigms and architectural choices aligned with emerging challenges. 3) Evolving research topics: Display of the annual distribution of trending topics in social media integrity research involving LLMs.

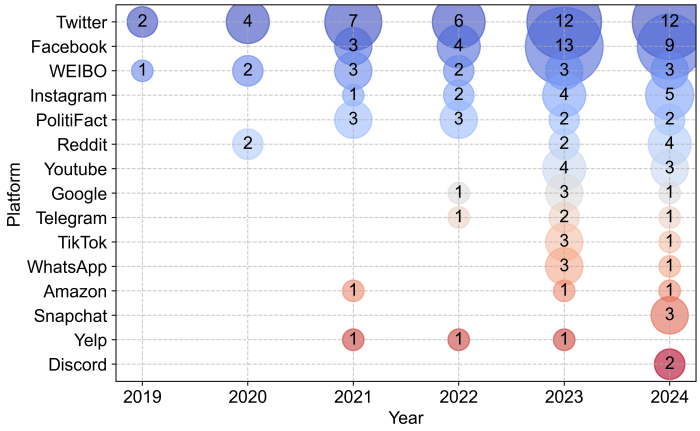

Figure 4. Temporal Analysis of LLM-based Social Media Integrity Research (2019-2024) Platform-specific research distribution visualized through bubble charts, where bubble size indicates study frequency. The visualization reveals temporal trends and platform-specific focus in LLM-driven integrity research.

versatility in both generating and analyzing content. Encoder-only models (BERT, RoBERTa) maintain their presence possibly due to their efficiency in classification and detection tasks. Additionally, encoder-decoder architectures (e.g., T5, BART) offer a hybrid approach that combines the strengths of both encoder and decoder structures, making them well-suited for tasks that require both deep understanding and fluent generation. This diversification in architectural choices suggests a growing recognition that different integrity challenges may require specialized architectural approaches.

Research Focus Distribution: Our temporal analysis of research themes (Figure 3) demonstrates significant shifts in research priorities. While traditional concerns like misinformation detection and content moderation remain fundamental, recent years have witnessed increased attention to emerging challenges. One obvious trend is the surge in research addressing sophisticated bot detection, multimodal deepfake analysis, and privacy-preserving integrity solutions. This evolution reflects both the increasing complexity of integrity threats and the research community's proactive response to emerging challenges.

Platform Distribution: The bubble chart visualization (Figure 4) illustrates the distribution of research across social media platforms. Twitter (now X) emerges as the primary research platform, largely

 doi:10.20944/preprints202508.1280.v1

due to its robust API infrastructure and rich dataset availability for studying automated behaviors and misinformation patterns. While established platforms like Facebook and Reddit maintain steady research presence, emerging platforms such as Discord and Mastodon represent growing research interests, particularly in studying integrity challenges in distributed social networks.

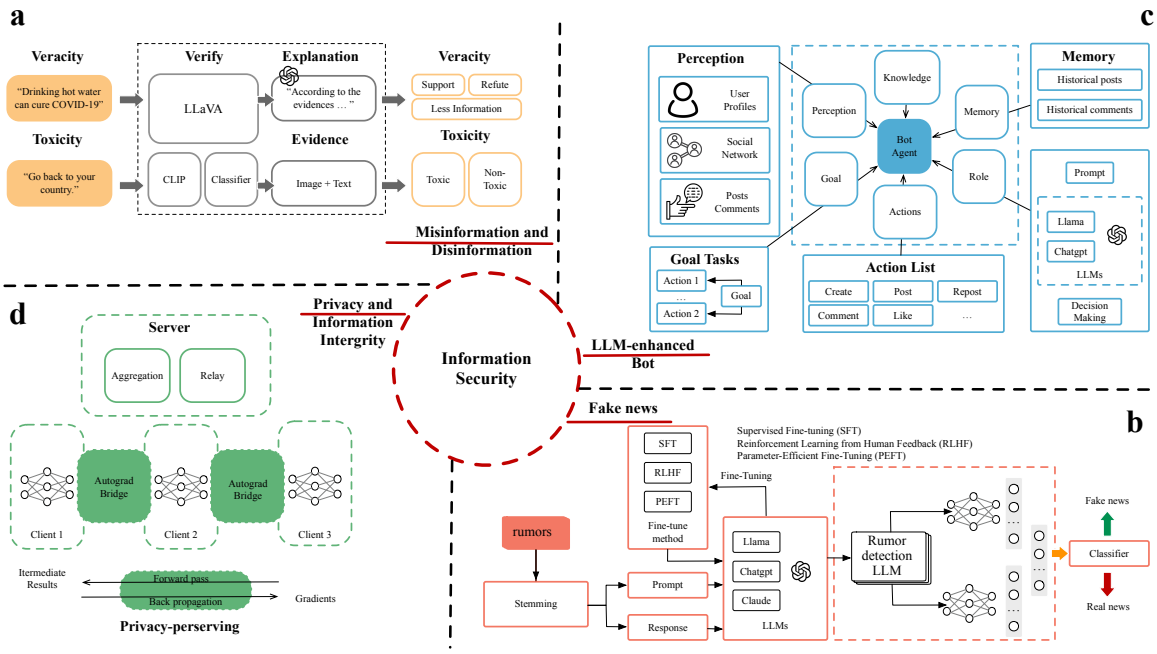

Figure 5. This figure illustrates four representative subdomain workflows (a–d) in information security research. (a)Misinformation and Disinformation: Lee et al. [86] propose a multimodal knowledge transfer approach using LLaVA and CLIP to improve fact-checking and rumor detection. (b)Fake News: Lai et al. [84] develops automated detection systems for identifying and combating synthetic fake news content. (c)LLM-enhanced Bot Detection: Liu et al. [94] demonstrates advanced techniques for identifying automated accounts and coordinated bot networks. (d)Privacy and Information Integrity: Su et al. [148] explores privacy-preserving mechanisms for protecting user data while maintaining information integrity in social media environments.

4. Research Outlook

4.1. Topic 1: Information Disorder

4.1.1. Misinformation and Disinformation

The proliferation of misinformation and disinformation on social media platforms represents one of the most pressing challenges in our digital age, with recent studies revealing that false information propagates as rapidly as truthful content [57,152]. Within this landscape, the advent of LLM marks a critical turning point, introducing a complex dynamic where these models serve both as potential solutions and possible sources of misinformation. While LLMs demonstrate remarkable capabilities in detecting patterns and inconsistencies across multiple languages and formats [43,163,179], their sophisticated generative abilities simultaneously present unprecedented challenges by producing highly convincing synthetic content [3,18,139,145,166]. This dual nature of LLMs in the information integrity landscape requires careful examination of both their potential benefits and risks in social media environments, particularly as automated accounts continue to amplify the spread of both authentic and false information.

 doi:10.20944/preprints202508.1280.v1

Table 2. Representative techniques of Misinformation.

Info-disorder Cases		**Description**	**References**
Unintentional Generation		Unintentional generation, known as hallucinations, occurs naturally through their generation properties.	[31,139]
Intentional Generation		The deliberate misuse of LLMs to create disinformation.	[16,124]
Techniques	**Methods**	**Description**	**References**
Zero-shot Detection	ChatGPT LLaMA	The ability to detect both human-written and LLM-generated misinformation.	[27]
Fine-tuned Detection	FACT-GPT FactLLaMA	Fine-tuned LLMs to enhance fact-checking to complement human expertise.	[28,30,86]
Knowledge-based Detection	LEMMA	Introducing external knowledge to the detection algorithm.	[86,164,166]

4.1.2. Misinformation and Disinformation Potentials: How LLMs Enhance Detection and Prevention.

LLMs' potential in misinformation detection has evolved from foundational approaches to sophisticated methods. As shown in Table 2, detection techniques can be broadly categorized based on their methodology, including early BERT-based models and more advanced fine-tuned, knowledge-based approaches. Early BERT-based methods demonstrated initial capabilities in misinformation detection tasks [2,43,71]. For example, Dulhanty et al. [43] explores automated methods through position detection, while Jiang et al. [71] develops BERT-based transfer-learning models.

Strong capability as the LLMs have, their performance on unseen data or circumstances, named as zero-shot setting, can be further improved [27]. Recent advances have produced several specialized detection strategies. Fine-tuned models like FACT-GPT [30], which leverages GPT-4 generated datasets for automated fact-checking, and FactLLaMA [28], which employs instruct-tuning for automated verification, represent significant progress in detection capabilities. Knowledge-based approaches have further enhanced detection systems, with ChatGPT being used to construct knowledge-based semantic structures [166]. Multimodal approaches have emerged through the integration of LLaVA and CLIP for knowledge transfer [86], while LEMMA [164] combines Large Vision Language Models with Chain-of-Thought reasoning for comprehensive detection. Recent frameworks like MUSE [181] attempt to address this through automated fact-checking using online credible sources, generating queries, and providing explanations with credible source links.

Advanced adaptive techniques have also been developed, such as MetaAdapt [172], which employs meta-learning for domain-adaptive few-shot detection, demonstrating superior performance over existing baselines. Zero-shot detection methods [27] have shown promise in identifying both human-written and LLM-generated misinformation, with performance varying based on content characteristics.

Detection performance is influenced by multiple factors, with content length being a key determinant—longer texts often yield higher accuracy due to richer contextual cues [27]. To address domain shifts and imbalanced data distributions, adaptive methods such as MetaAdapt have been developed [172]. In parallel, researchers have explored advanced prompting strategies and multi-step verification workflows [86,166], incorporating multimodal tools like CLIP and LLaVA to further enhance model robustness [86]. Despite these advances, significant challenges remain. Detection systems still struggle with subtle manipulations and short-form content, while accuracy varies based on text origin and model architecture. The rapid evolution of generative models necessitates continuous updates to detection mechanisms, as evidenced by systems like FACT-GPT [30] and FactLLaMA [28].

 doi:10.20944/preprints202508.1280.v1

4.1.3. Misinformation and Disinformation Challenges.

As shown in Table 2, LLM-generated misinformation presents two distinct categories of challenges. The first is unintentional generation or hallucinations [31,139], which often occur due to the intrinsic properties of LLMs. These hallucinations are particularly challenging, as they appear highly plausible in fine-grained details such as dates, names, and numbers, making them difficult to detect even when models attempt to generate factual content. Shah et al. [139] explore this dual role of LLMs in both the combating and the potential generating of misinformation, examining their impact on social media consumption and various domain applications. Christensen et al. [31] further demonstrate these challenges through specific cases like tourism planning.

The second category involves intentional generation - the deliberate misuse of LLMs for disinformation [16,124]. This ranges from minor alterations to critical changes that dramatically alter meaning. Radivojevic et al. [124] demonstrate through experiments with GPT-4, LLama2, and Claude that humans can only identify the origin of such content 42% of the time, with persona's influence exceeding human perception. Barman et al. [16] further explore LLMs' potential in generating multimedia disinformation, emphasizing the need for ethical oversight and multi-stakeholder collaboration. These challenges are compounded by quality control issues, where verification must evaluate multiple aspects, including factual consistency and contextual alignment. The integration of these challenges with social media platforms creates additional complexities [43,71,166]. The scalability and automation capabilities of LLMs enable rapid generation of false content that can be automatically adapted across platforms. This technological capability, combined with sophisticated audience-targeting and platform-specific optimization, significantly enhances the potential for echo chamber reinforcement. These issues are further complicated by biased datasets, adversarial attacks, and the risk of users taking automated fact-checking outputs as definitive truth without proper judgment [43,71]. Additionally, as noted by Yang et al. [166], data bias and topical bias in polarized comments remain underexplored challenges in current research.

 doi:10.20944/preprints202508.1280.v1

Table 3. Comparison between Fake News and Verified Facts.

✗ Fake News	✓ Factual Correction
✗ Starbucks is sponsoring the Republican National Convention in Milwaukee.	✓ Social media users are claiming that Starbucks, known for taking strong positions in support of progressive political issues, is sponsoring the RNC.
✗ A bill signed into law this week by Michigan Gov. Gretchen Whitmer prohibits vote recounts based on election fraud allegations.	✓ The bill, SB 603, does not prohibit such recounts, according to two state senators involved in updating the laws around recounts. It stipulates that candidates may request a recount if they have a "good-faith belief" that they would have had a "reasonable chance" to win the election if not for an "error" in the vote-counting process. That means that the number of votes the petitioning candidate requests to be recounted must be greater than the difference of votes between them and the winning candidate.
✗ A video shows President Joe Biden trying to sit in a chair that wasn't there during a ceremony in Normandy, France, commemorating the 80th anniversary of D-Day.	✓ The video, in which Biden's chair is for the most part clearly visible, is cut before the president sits down. Full footage of the ceremony shows the president looking over his shoulder for his chair and pausing before taking a seat.
✗ A video shows a worker at a voter registration drive in Florida registering people to vote without asking for proof of citizenship.	✓ Shortly before the U.S. House of Representatives on Wednesday passed a bill requiring proof of citizenship to register to vote, social media users shared a video of a voter registration drive in Palm Beach, Florida, to raise questions about noncitizens voting in U.S. elections.

The "AP Fact Check" section [10] is a specialized platform by AP dedicated to verifying and clarifying rumors and fake news circulating online. Operated by an experienced team of journalists, it follows a rigorous process of multi-source verification, public records examination, and subject confirmation to provide readers with authoritative, transparent corrections and explanations. This table contains the information we retrieved from that URL https://apnews.com/ap-fact-check.

Table 4. Representative techniques of Fake News Detection.

Techniques	Model	Application	References
Transformer Learning	BERT, RoBERTa, XLM-RoBERTa	Text-based detection; rumor analysis; cross-lingual detection	[35,37,97] [26,55,101]
Enhance Explainability	DistilBERT +SHAP	Enhancing model interpretability; explainable decision-making	[135] [99]
Multi-modal Fusion	BDANN, EANBS, FND-CLIP, ChatGPT	Joint analysis of textual and visual information	[137,176] [46,100]
Prompt-based Learning	ChatGPT, GenFEND	Zero-shot classification; training data augmentation;debiasing	[4] [66,113]
Few-shot Learning	RumorLLM	Reducing annotation costs and improving performance in low -resource settings	[84]
Privacy-preserving Learning	AugFake-BERT	Collaborative training without data sharing, ensuring data privacy	[78]
Domain Transfer Learning	CT-BERT RoBERTa	Domain-adaptive or region-specific misinformation detection	[101] [147]

4.1.4. Fake News Potentials: LLMs as Detection Tools

Table 4 presents a systematic overview of fake news detection techniques, illustrating their evolution from basic transformer models to sophisticated multi-modal and domain-specific applications.

 doi:10.20944/preprints202508.1280.v1

Among these techniques, transformer-based learning forms the foundation of modern fake news detection, with models like BERT and RoBERTa demonstrating strong performance in text analysis. First, basic models like BERT, RoBERTa, and XLM-RoBERTa excel in text-based detection, with fine-tuned RoBERTa models specifically targeting COVID-19 misinformation [26,101]. Second, cross-lingual detection capabilities are enhanced through XLM-RoBERTa-CNN combinations [55]. Third, hybrid models like BDANN integrate BERT with VGG-19 for initial multi-modal detection [176], while DistilBERT improves result explainability through SHAP techniques [13].

Advanced multi-modal detection has evolved through three key innovations. First, FND-CLIP and EANBS process text and visual inputs simultaneously through contrastive learning and integrated BERT-CNN architectures [46,137], enabling the detection of inconsistencies between textual claims and associated images. Second, GPT-4 enhances detection by identifying subtle cross-modal inconsistencies [100], particularly effective in cases where text and images have been manipulated to create misleading narratives. The model's strong performance in cross-modal reasoning helps identify cases where images have been repurposed with false textual contexts or where subtle visual manipulations contradict textual claims. Additionally, generative frameworks have advanced detection capabilities in two ways: ChatGPT enables zero-shot classification in low-resource settings [4], allowing effective detection even without extensive training data, while GenFEND enriches training data through scenario simulation [113], creating diverse synthetic examples of misinformation patterns. These approaches are particularly valuable for emerging topics where labeled training data is scarce, helping detection systems remain effective against novel forms of misinformation.

Domain Transfer and Cross-lingual Learning approaches address the challenge of adapting detection models across different domains and languages. CT-BERT demonstrates effective domain adaptation for COVID-19 misinformation [101], while RoBERTa-based models show promising results in cross-lingual transfer for Romanian political misinformation detection [147]. These domain transfer techniques are particularly crucial when dealing with emerging topics or low-resource languages. For instance, ChatGPT-based transfer learning has shown effectiveness in adapting bias detection across different cultural contexts [66]. Meanwhile, complementary techniques like active learning in RumorLLM [84] and federated learning in AugFake-BERT [78] enhance the efficiency and privacy of these domain adaptation processes. Beyond detection, LLMs advance fake news understanding through simulation and analysis. The Fake News Propagation Simulation framework studies diffusion patterns and virality factors [95], while Fact Agent models analyze dissemination patterns and develop real-time mitigation strategies [90].

4.1.5. Fake News Emerging Challenges

Social media platforms have fundamentally transformed fake news propagation, enabling rapid and extensive reach with minimal resources [78,123,157,176]. This accelerated dissemination significantly erodes public trust in both social media platforms and institutions [72,137,171]. As illustrated in Table 3, real-world examples of fake news reveal how easily misinformation can be framed to appear credible, emphasizing the need for timely and accurate detection mechanisms. The challenge is particularly acute in low-resource settings and across linguistic barriers. For example, Rathinapriya et al. [128] reveal substantial accuracy drops when detection systems trained on high-resource languages were applied to regional Indian languages, emphasizing the critical gap in cross-lingual detection capabilities. The speed and scale of social media sharing make traditional fact-checking methods increasingly inadequate, as false information can reach massive audiences before corrections can be implemented [98].

The increasing prevalence of multimedia content presents significant technical challenges [4,46]. Current detection frameworks face multiple limitations: they rely heavily on human annotation, limiting scalability [56]; they struggle with cross-platform content verification; and they often fail to capture context-dependent nuances. Additionally, generative AI advancements enable the creation of increasingly sophisticated and credible false content [91], making detection more complex. These

 doi:10.20944/preprints202508.1280.v1

technical challenges are particularly evident in real-time detection scenarios, where the need for immediate response conflicts with the requirement for accurate verification.

The societal impact of fake news manifests across multiple critical domains. In healthcare, misinformation during the COVID-19 pandemic led to public confusion, vaccine hesitancy, and inadequate crisis responses [13,26,55]. In politics, fake news serves as a powerful tool for manipulation and polarization [147,161], undermining democratic processes and public discourse. In economics, it disrupts markets and consumer behavior through false financial information and manipulated market signals [62,84]. These problems are particularly severe in regions with limited media literacy [98,128,131], where both active and passive consumers [113] can be significantly influenced by false narratives. Fake news exploits emotional and cultural elements to enhance its credibility and impact [46,127,161]. Its sophisticated adaptation to regional contexts and local sensitivities maximizes societal impact [98,128], particularly through multimedia elements that increase persuasiveness [4,46]. The emotional manipulation often targets existing societal divisions, amplifying conflicts and polarization. These challenges are further intensified by social media's connectivity [78,123,157,176], where echo chambers and algorithmic content promotion can rapidly spread misinformation within susceptible communities, significantly affecting public discourse and social cohesion.

4.2. Topic 2: Social Bot

4.2.1. Potential of LLM-Enhanced Bots in Information Integrity

LLM-enhanced bots demonstrate significant potential in maintaining and enhancing information integrity on social media platforms. These advanced bots serve as powerful tools for automated content verification, real-time fact-checking, and proactive misinformation prevention through their sophisticated pattern recognition and contextual analysis capabilities Panagiotou et al. [115], Xu et al. [162]. Their multilingual abilities enable cross-language content verification [51,107], while their advanced natural language processing capabilities allow for more accurate identification of suspicious patterns in both user behavior and content distribution [52,54]. The integration of LLMs has particularly enhanced bots' ability to understand nuanced context, engage in interactive fact-checking, and provide immediate feedback with educational resources about potential misinformation [63,138]. Emerging applications suggest a promising future where LLM-enhanced bots can participate in collaborative verification systems, implement adaptive learning mechanisms for improved detection accuracy, and deploy context-aware intervention systems for maintaining the health and reliability of social media information ecosystems [122,162].

Research into using LLM-enhanced bots for maintaining information integrity has evolved significantly, with substantial advances in detection and prevention capabilities. Early detection systems established fundamental approaches that continue to influence current solutions. Dukić et al. [42] and Heidari et al. [65] pioneered the integration of emoji embeddings and sentiment features with BERT models, demonstrating significant improvements in bot detection accuracy. These foundational works were extended by Xu et al. [162] through the combination of ALBERT with Bi-LSTM and self-attention mechanisms, establishing new benchmarks in bot spam detection performance. Multilingual detection capabilities have seen remarkable advancement. A series of studies demonstrated effective cross-lingual approaches. Panda et al. [116] developed models for detecting misinformation across English, Bulgarian, and Arabic; Garcia-Diaz et al. [51] achieved over 97% accuracy in Spanish language satire detection; and Milkova et al. [107] successfully detected value-expressive posts in Russian social media. These works collectively demonstrate the adaptability of LLM-based bot detection systems across different linguistic contexts. Advanced architectural approaches have significantly improved detection capabilities. Kumar et al. [83] and Guo et al. [59] developed hybrid frameworks combining BERT with graph convolutional networks, while Garcia-Silva et al. [52] demonstrated the superiority of generative transformers in bot detection tasks. Recent work by Ghanem et al. [54] and Raga et al. [125] has pushed performance boundaries further, achieving accuracy rates above 98% across different platforms.

Real-time monitoring and prevention bot systems have emerged as crucial defensive tools. Panagiotou et al. [115] developed News Monitor, achieving high accuracy in rumor identification, while

 doi:10.20944/preprints202508.1280.v1

Kawintiranon et al. [76] and Catelli et al. [23] demonstrated effective approaches for context-specific bot spam detection. These systems have been enhanced by recent work from Ramamonjisoa et al. [126] and Bonechi et al. [17], who developed sophisticated automated moderation systems for bots. Specialized detection approaches have shown promising results in specific domains. Jones et al. [73] and Sallah et al. [133] developed bot systems specifically targeting child protection on social media, while Riyantoko et al. [129] demonstrated effective SMS bot spam classification using combined LSTM and BERT approaches. These specialized applications demonstrate the adaptability of LLM-based detection bot systems to specific security challenges.

Table 5. Representative techniques of LLM-enhanced bot detection.

Techniques	Methods	Description	References
Hybrid neural Detection	BERT+GCN, ALBERT+Bi-LSTM, Transformer method	Advanced bot detection using hybrid neural architectures	[83,162] [54,59]
Multilingual Processing	XLM-RoBERTa, mBERT, Multilingual GPT	Cross-lingual bot detection and content analysis across different languages and cultural contexts	[51,116] [107]
Real-time Monitoring	News Monitor, Moderation Systems	Continuous monitoring and instant detection of malicious activities	[115,126] [17]
Domain-specific Defense	Child Protect Systems, Financial Monitors	Specialized bot detection systems for specific sectors or use cases	[73,133] [129]
Multimodal Analysis	GPT-4V, CLIP-based Models	Joint analysis of text, images, and other media types for comprehensive bot detection	[100,104]
Privacy Preserving	Federated Learning, Differential Privacy	Collaborative bot detection while maintaining data privacy	[121]
Proactive Prevention	Sheaf Theory + LLMs, Topic Expansion	Early warning and prevention systems before bot attacks occur	[69,94]
Context-aware Detection	SBERT + Topic Models Contextual Embeddings	Understanding and analyzing bot behavior in specific contexts	[9,94]
Hybrid Defense	LSTM+BERT, Ensemble Methods	Integration of multiple detection techniques for robust defense	[6,134]

4.2.2. Challenges and Risks of LLM-Enhanced Bots

The evolution of LLM-enhanced bots and their impact on information integrity has been a growing concern since 2019. Early research by Gupta et al. [60] and Ghanem et al. [53] revealed how malicious bots were increasingly adopting transformer-based architectures to generate more convincing spam content on Twitter. While their detection models achieved accuracy rates above 98% for traditional bot behaviors, these studies also pointed out an emerging challenge: as bots incorporated more sophisticated LLM capabilities, they became increasingly difficult to distinguish from human users. Recent studies by Li et al. [87] and Yang et al. [167] further revealed how these bots evolved to form large-scale coordinated networks, manipulating information across various domains, from health misinformation to financial markets [150]. The COVID-19 pandemic marked a critical period in the evolution of malicious bot capabilities. Multiple research teams documented the spread of health-related misinformation. Sai et al. [132] and Kar et al. [75] demonstrated how LLM-enhanced bots effectively disseminated COVID-19 misinformation across multiple languages. This was further confirmed by Kim et al. [80] and Sharma et al. [141], who found these bots achieved high accuracy in mimicking legitimate health information sources. Pranto et al. [118] specifically identified ten distinct topics in fake news, illustrating the sophisticated nature of cross-lingual misinformation campaigns.

Recent advances in multimodal capabilities have further complicated the threat landscape. Lyu et al. [100] show how GPT-4V-powered bots can effectively manipulate both textual and visual content.

 doi:10.20944/preprints202508.1280.v1

This capability was further explored by Mehta et al. [104], who demonstrated how these bots form sophisticated user communities to amplify their impact. The integration of advanced language models has enabled these bots to generate highly personalized phishing content, as documented by Haynes et al. [63] and Atzori et al. [11], who showed how bots exploit vulnerabilities across multiple social networks. The cross-lingual capabilities of these bots present particular challenges. Studies by Shukla et al. [143] and Ahmed et al. [1] revealed how LLM-enhanced bots effectively adapt their content across different languages and cultural contexts. This adaptability was further demonstrated by Harrag et al. [61] in their analysis of Arabic language manipulation, achieving deception rates of up to 98%.

The development of defense mechanisms against LLM-enhanced bots has evolved into a multi-layered approach, incorporating various technological and methodological innovations. Advanced Detection Architectures have emerged as a primary defense strategy. Alshattnawi et al. [6] demonstrated significant improvements in spam detection accuracy through deep neural networks with contextualized word embeddings, consistently achieving 10-15% improvement over traditional approaches. This work was complemented by Sangher et al. [134], who integrated LSTM and BERT-based transformers to identify sophisticated cybercrime patterns. The effectiveness of these approaches was further validated by Shafee et al. [138], who evaluated various LLM chatbots for cybersecurity threat awareness.

Proactive prevention systems represent another crucial defensive layer. Huntsman et al. [69] proposed innovative approaches using LLMs and sheaf theory to detect textual inconsistencies, while Milner et al. [108] developed lightweight phishing detection algorithms specifically optimized for mobile devices. These systems have been enhanced by Puppala et al. [121], who introduced a Federated Learning-based GPT system that ensures privacy while maintaining bot detection effectiveness. Cross-platform and Multilingual Defense strategies have become increasingly important. Liu et al. [94] developed early identification methods using topic expansion and SBERT models, while Askari et al. [9] conducted field experiments to evaluate the effectiveness of bot-based interventions in promoting legitimate news consumption. These approaches demonstrate the importance of comprehensive, platform-agnostic defense mechanisms.

4.3. Topic 3: Privacy

The integration of LLMs into social media platforms has fundamentally reshaped the privacy landscape. With their ability to infer and memorize personal details across modalities, LLMs introduce privacy risks far beyond traditional NLP systems. In this section, we examine these emerging privacy challenges, where Table 6 summarizes current privacy-preserving techniques and their applications in social media contexts. These privacy concerns and protective measures reflect the ongoing evolution of social media information integrity in LLM-enhanced environments.

Table 6. LLM Applications and Their Privacy Implications: A Comprehensive Overview of Privacy-preserving Techniques in Social Media Contexts.

Techniques	Methods	Description	References
Knowledge Distillation	Privacy-aware model compression and transfer	Minimize sensitive data exposure during training	[49,180]
Differential Privacy	DP-SGD and privacy-preserving optimization	Formal privacy guarantees for model training	[67,89]
Federated Learning	Decentralized model updates and training	Privacy-preserving distributed learning	[148,169]
Adversarial Protection	Text perturbation and obfuscation techniques	Prevent unauthorized inference and linking	[39,149]
Privacy-Preserving Attention	Modified attention mechanisms for privacy	Secure information processing in LLMs	[22,117]

 doi:10.20944/preprints202508.1280.v1

The challenge of protecting sensitive data while maintaining model utility presents a fundamental tension in LLM deployments. Cai et al. [21] identify critical challenges in balancing analytical capabilities with privacy protection, particularly in mental health discussions. Traditional privacy-preserving techniques prove increasingly inadequate, as Patsakis and Lykousas [117] and Mattern et al. [103] show how LLMs can still infer user traits through subtle linguistic cues even after standard de-identification. This has led to exploration of more robust approaches, with Huang et al. [67], Li et al. [89], Wiseman et al. [160] investigating differential privacy strategies, while Su et al. [148] explores federated learning solutions for privacy-aware model updates. As Martin et al. [102] and Dickson and McCauley [38] emphasize, there is an urgent need to develop stronger ethical guidelines and resolve the fundamental tension between analytical utility and privacy preservation.

4.3.1. Potential in Privacy-Preserving Techniques to Guarantee Information Integrity

Privacy-preserving techniques for LLM-enhanced environments show promising potential in addressing current privacy challenges. Knowledge distillation methods proposed by Zhao and Caragea [180] demonstrate how label prediction can effectively abstract content while minimizing sensitive data exposure. In URL filtering contexts, P-BERT introduced by Supriya and Akki [149] shows significant potential in integrating deep feature extraction with BERT for malicious link detection while reducing leakage risks. These advancements directly address the limitations identified by Patsakis and Lykousas [117] and Mattern et al. [103], where traditional anonymization methods fail against LLMs' ability to infer user attributes from residual linguistic signals.

Advanced privacy protection mechanisms are emerging to tackle the issue of raw data exposure during processing. Differential privacy approaches show particular promise, with Coffey et al. [33] demonstrating how differentially private stochastic gradient descent during fine-tuning can effectively limit model memorization of individual data points. Privacy-preserving synthetic data generation, such as SynthPAI by Yukhymenko et al. [173], offers a solution to avoid direct handling of personal information. Additionally, language-specific approaches for PII protection, as demonstrated by Jang et al. [70] for Korean data, provide promising templates for extending privacy safeguards across different linguistic contexts.

Emerging solutions for deep anonymization are addressing the limitations of conventional de-identification methods. Linguistic steganography techniques Coffey et al. [33] show potential in encoding sensitive content in forms only intelligible to intended recipients, while authorship obfuscation frameworks Bao and Carpuat [14] demonstrate how reinforcement learning can effectively conceal personal stylistic features without compromising meaning. These approaches directly address the challenges identified by Patsakis and Lykousas [117] and Mattern et al. [103] regarding LLMs' ability to infer user traits through subtle linguistic cues.

Real-time monitoring and accountability systems represent crucial advancements in privacy protection. For example, Kim et al. [81] and Asimopoulos et al. [8] demonstrate potential in proactively detecting and suppressing sensitive information in model outputs. Federated learning approaches by Yao et al. [169,170] show promise in reducing centralized data storage vulnerabilities, while explainable privacy frameworks proposed by Mireshghallah et al. [109] offer solutions for transparency and accountability concerns. These developments directly address the need for context-aware privacy reasoning in diverse and dynamic social media interactions, providing comprehensive solutions for protecting user data while maintaining system functionality.

In recent years, significant advances have been made in addressing LLM privacy challenges, with a focus on several prominent research directions. First, model inversion attacks [22,50,112,158] exploit output probabilities to reconstruct hidden prompts or input data, posing serious threats in LLM-as-a-service settings where prompts may encode private instructions. In terms of data leakage risks, studies have revealed two major vulnerabilities: LLMs can retain and reproduce sensitive information even when not overfitted [15,41,74,136], and through membership inference attacks, adversaries can determine whether specific user data was included in the training set using techniques like SPV-MIA with paraphrasing and self-calibrated reference models [29,120,174,175]. To address these

 doi:10.20944/preprints202508.1280.v1

challenges, researchers have developed various privacy-preserving approaches: differential privacy techniques [67,89,160] implement DP-SGD to provide formal privacy guarantees, while federated learning solutions [148] explore decentralized model adaptation to prevent centralized data exposure, though both approaches must balance privacy protection with model utility. These privacy vulnerabilities and protective measures represent the current landscape of privacy challenges in LLM-enhanced social media environments.

4.3.2. Privacy Challenges Arising from LLM-Induced Information Integrity Collapse

The integration of LLMs in social media applications presents critical privacy challenges that demand immediate attention. The primary challenge lies in preventing unauthorized information inference: Staab et al. [146] reveal how LLMs can breach privacy by deducing sensitive personal information from pseudonymized content through subtle linguistic markers and reconstructing detailed user profiles by connecting seemingly unrelated pieces of information. The cross-platform privacy challenge is particularly pressing, as Treves et al. [153] exposes how tools like RURLMAN can breach user privacy by automatically linking identities across multiple platforms through shared URLs, creating comprehensive digital footprints without user consent. This vulnerability is further amplified by the challenge of preventing synthetic identity abuse, where Ayoobi et al. [12] identifies how these models enable the creation of deceptive profiles on professional networks like LinkedIn for malicious purposes. As Dogan et al. [39] emphasizes, there is a critical need to develop robust safeguards against the systematic linking of online personas with real-world identities, fundamentally challenging traditional privacy protection approaches.

Real-time deployment of LLMs introduces additional urgent privacy challenges that require innovative solutions. Brown et al. [20] demonstrates the pressing need to establish clear boundaries between private and public information in ChatGPT's live platform interactions, while Dou et al. [40] underscores the challenge of preventing unauthorized behavior pattern analysis that could expose users' personal lives, health, and relationship information. In sensitive topic analysis, Cai et al. [21] identifies the critical challenge of balancing analytical capabilities with privacy protection, particularly in mental health discussions where inadvertent exposure of personal insights poses significant risks. These challenges necessitate the immediate development of enhanced privacy frameworks, as Martin et al. [102] argues for stronger ethical guidelines in governing LLM deployment, and Dickson and McCauley [38] emphasizes the urgent need to resolve the fundamental tension between analytical utility and privacy preservation in these sensitive contexts.

5. Landscape Analysis: Potentials and Challenges

5.1. Landscape: Opportunities in Social Media Information Integrity

As detailed in Sections 3-5, LLMs have delivered multi-dimensional gains for social-media information integrity, as illustrated in Figure 6. When enhanced through domain-specific fine-tuning or retrieval-augmented generation (RAG), they achieve significant improvements in multilingual misinformation screening recall, reduce fact-checking latency through automated source verification, identify covert bot networks through pattern analysis, and generate privacy-aware prompts that strengthen user media-literacy. At the same time, careful system design, such as contextual-integrity guards, federated fine-tuning, and cost-aware routing, can keep these advances affordable and policy-compliant. In what follows, we summarize the landscape of LLM opportunities across key dimensions of social media information integrity.

 doi:10.20944/preprints202508.1280.v1

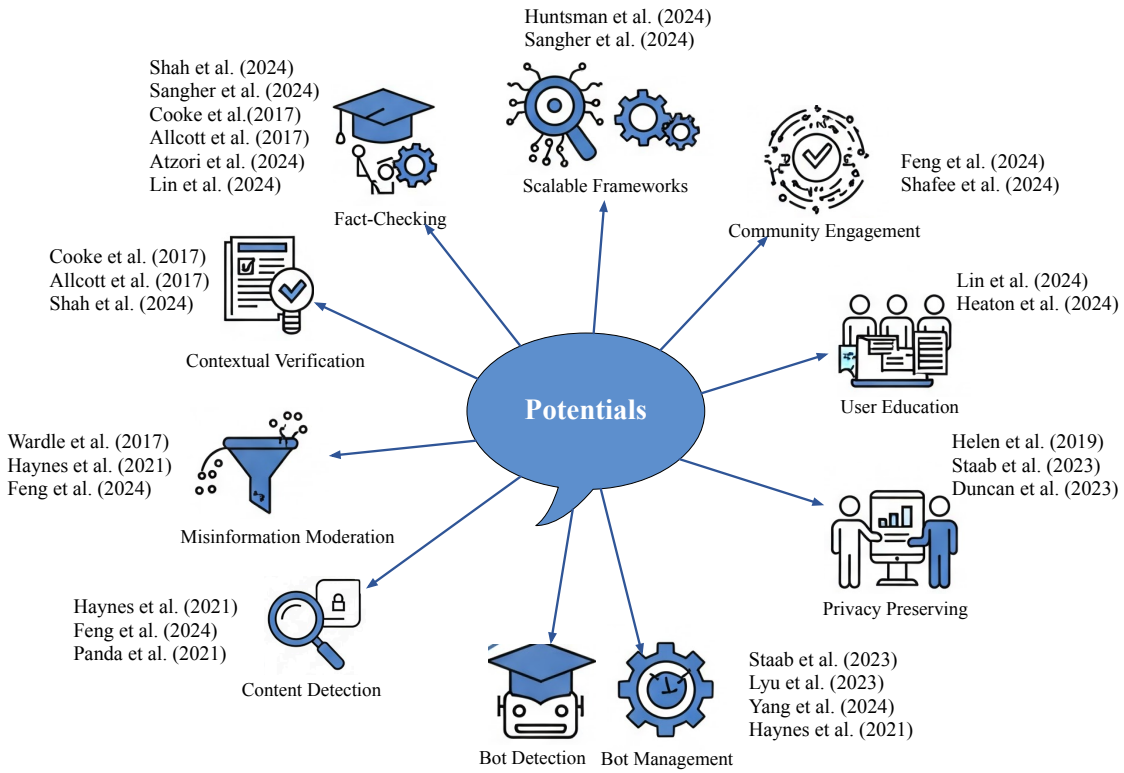

Figure 6. Potentials in LLMs for social media information integrity. The figure illustrates key capabilities of LLMs in enhancing content verification, automated fact-checking, and content moderation, while also illustrating advances in privacy-preserving processing, user education, community engagement, and cost-effective deployment frameworks. Each directional component references recent research (e.g., Wardle et al. (2017) - [159], Haynes et al. (2021) - [63], Panda et al. (2021) - [116], Cooke et al. (2017) - [36], Allcott et al. (2017) - [5], Shah et al. (2024) - [139], Atzori et al. (2024) - [11], Lyu et al. (2023) - [100], Yang et al. (2024) - [167], Duncan et al. (2023) - [44], Puppala et al. (2024) - [121], Nissenbaum et al. (2019) - [114], Lin et al. (2024) - [93], Heaton et al. (2024) - [64], Staab et al. (2023) - [146], Feng et al. (2024) - [48], Shafee et al. (2024) - [138], Panagiotou et al. (2021) - [115], Huntsman et al. (2024) - [69], Sangher et al. (2024) - [134])

Detect & Moderate Misinformation.

LLMs now raise recall by 6–10% on the HARMONY-22 hate-and-misinformation benchmark when fine-tuned with contrastive prompts, and they achieve comparable gains on cross-language rumour datasets [82,85]. By modeling narrative intent, they distinguish inadvertent misinformation from deliberate disinformation, extending Wardle's conceptual split [159]. This capability allows the same model to recognize both careless sharing and coordinated campaigns in diverse linguistic and cultural settings, even in low-resource languages [116]. Recent audits confirm that LLMs can detect nuanced health and political falsehoods that evade traditional filters [48,63].

Multimodal Fact-Checking & Contextual Verification.

Coupling GPT-4 claim decomposition with retrieval-augmented generation shortens fact-checker latency by roughly 35% on the FFRR benchmark [130], while agentic pipelines like FactAgent decompose verification into micro-tasks that crowd workers validate asynchronously [48]. Cross-platform context analysis uncovers temporal or source inconsistencies that headline-only systems overlook [139], and when these pipelines incorporate news format, presentation style, and narrative evolution tracking, they can flag emerging misinformation trends earlier in the news cycle [134]. Recent studies report macro-F1 scores around 0.82 on real-time news streams [55], surpassing traditional fact-checking workflows [5,11,36,93].

 doi:10.20944/preprints202508.1280.v1

Bot Detection, Analysis & Management.

LLM-enhanced detectors jointly model textual semantics, image cues, temporal "digital-DNA" traces, and interaction graphs, yielding up to 9 percentage point gains in F1 score on the TwiBot-22 benchmark [47,100,146]. By analyzing behavioral patterns, content-generation fingerprints, and network dynamics, they surface coordinated, cross-platform inauthentic activity [5,167]. These systems can also separate sophisticated social-engineering bots from legitimate automation [63], enabling fine-grained intervention policies that uphold overall platform integrity.

Privacy-Preserving Content Processing.

Grounded in Nissenbaum's contextual-integrity theory [114], modern LLM pipelines can deliver personalised curation while respecting user privacy. Audits such as LLM-CI reveal that vanilla prompts leak context-specific private attributes in 39% of scenarios; embedding contextual-integrity constraints cuts that leakage by roughly 40% without sacrificing utility [5]. Complementary safeguards include differential-privacy masking and other privacy-preserving analytics [146], transparent data-handling and audit trails [44], and consent-aware model updates via federated or split learning [121]. Together these measures enable platforms to offer fine-grained, trustworthy recommendations while upholding robust privacy guarantees.

User Education & Media-Literacy Support.

Instruction-tuned LLMs can generate adaptive "trust cues," critical-thinking scaffolds, and media-literacy tools that raise users' misinformation-recognition accuracy by about 12% in controlled studies [64]. Beyond static tips, the models dynamically tailor explanations and verification checklists to a user's demonstrated knowledge and engagement level, drawing on multilingual value-alignment work that transfers across languages with little performance loss [68]. They also incorporate practical guidance for source checking and claim verification [93], thereby empowering users to recognize manipulation techniques and exercise informed skepticism.

Community Engagement & Collaborative Verification.

LLMs catalyze constructive dialogue by generating prompts and replies that nudge users toward evidence-based discussion [146]. In civic fact-checking forums, such prompts have doubled the volume of user-submitted evidence while avoiding alert-fatigue patterns [55]. Agentic pipelines like FactAgent decompose verification into micro-tasks that crowd workers validate asynchronously [48], yet still manage resource constraints [138] and platform-integration complexity [126]. Additional mechanisms for monitoring participation and throttling notifications mitigate fatigue and sustain long-term engagement [115].

 doi:10.20944/preprints202508.1280.v1

5.2. Landscape: Challenges in Social Media Information Integrity

Figure 7. Challenges in LLMs for social media information integrity. The figure illustrates the key capabilities of LLMs in privacy vulnerabilities, cross-platform security risks, and fairness, as well as accountability, user rights, social impact, resource and deployment, performance bottlenecks, detection evasion, coordinated bots, and cross-lingual limitations. Each directional component references recent research (e.g., Feng et al. (2024) – [48], Sangher et al. (2024) – [134], Panda et al. (2021) – [116], Shah et al. (2024) – [139], Haynes et al. (2021) – [63], Staab et al. (2023) – [146], Yang et al. (2024) – [167], Atzori et al. (2024) – [11], Lin et al. (2024) – [93], Duncan et al. (2023) – [44], Heaton et al. (2024) – [64], Puppala et al. (2024) – [121], Shafee et al. (2024) – [138], Ramamonjisoa et al. (2024) – [126], Panagiotou et al. (2021) – [115], Huntsman et al. (2024) – [69].)

Our landscape analysis also reveals four primary challenge domains in LLM-based social media information integrity, as illustrated in Figure 7. These findings emphasize critical concerns across technical implementation, ethical considerations, deployment feasibility, and user interaction.

Detection and Verification Challenges

Technical obstacles in LLM-based information integrity now center on model-specific detection evasion, inference performance bottlenecks, and hallucination issues. Detection evasion remains a foremost obstacle: adversaries continually refine prompt engineering tactics to bypass LLM integrity filters, with evasion success rates reported as high as 29.6% [48]. Performance bottlenecks in LLM inference and prompt processing continue to hinder timely and accurate integrity verification across high-volume social media streams [48,134]. LLM cross-lingual capabilities remain limited, particularly in detecting misinformation across non-English or low-resource languages [116]. Most critically, model hallucination undermines trust in LLM-generated content, as spurious outputs can evade verification protocols and propagate misinformation [139].

Resource and Deployment Challenges

LLM deployment faces significant operational hurdles in resource management and system integration. The substantial computational demands of large language models and their complex integration requirements impede seamless deployment and maintenance of integrity solutions at scale [69,126,138]. The challenge is compounded by LLM-specific issues such as prompt optimization,

 doi:10.20944/preprints202508.1280.v1

model quantization needs, and the complexity of maintaining multiple model versions for different integrity tasks [115,126].

Privacy and Security Challenges

LLM-specific security risks center on prompt injection vulnerabilities, training data privacy, and model extraction threats. Privacy concerns are particularly acute with LLMs, as these models can potentially memorize and leak sensitive training data, while also being vulnerable to advanced prompt engineering exploits [63,146]. Additionally, adversaries can leverage LLMs to generate sophisticated phishing content and automate social engineering attacks at scale [167]. The integration of LLMs across multiple platforms further complicates security, as attackers can exploit variations in model behavior and security implementations across different services [11].

Fairness and Accountability Challenges

LLM-specific ethical considerations focus on model bias, output attribution, and societal impact. Fairness challenges arise from LLMs' training data biases and their potential to perpetuate or amplify societal prejudices in generated content [44,93]. The black-box nature of large language models complicates accountability, requiring new approaches to model interpretability and output verification. Additionally, the deployment of LLMs in social media integrity systems raises concerns about algorithmic transparency and the need for explicit user consent in content moderation [64,121].

6. Discussion

6.1. Key Findings

Leading Models and Research Paradigms. Based on our review, the field of LLM-driven social media information integrity has crystallized around several influential models and research approaches. First, in fact-checking and misinformation detection, LLM-based models like FACT-GPT [30] and FactLLaMA [28] have emerged as industry standards. These models excel at automated verification but face a critical challenge: they are prone to hallucinations in zero-shot and cross-domain scenarios, particularly when verifying specific details like dates, numbers, or named entities [31,139]. In the privacy and security domain, frameworks like ProPILE [81] and P-BERT [149], as well as federated learning approaches such as SocFedGPT [122], have emerged as leading solutions. While these LLM-based models have advanced detection capabilities, they also expose critical challenges in generalizability, interpretability, and robustness.

Trends and Future Breakthroughs. The research frontier is moving toward multimodal, cross-model ensemble, and human-in-the-loop systems. Multimodal detection frameworks that integrate text, metadata, and contextual signals have surpassed traditional text-only methods by 5–10% in accuracy [96], while ensemble verification systems achieve 8–12% higher detection rates for subtle misinformation. In privacy protection, differential privacy [89] and linguistic steganography [14] provide both theoretical and practical advances.

Detection systems are evolving from single-model approaches to multimodal, ensemble, and explainable frameworks. By fusing text, images, and social network structures, multimodal systems can significantly improve the detection of complex and adversarial misinformation. Ensemble methods (e.g., model stacking, expert systems) can reduce individual model bias and enhance robustness against adversarial content. At the same time, explainability mechanisms provide transparent rationales for moderation decisions, fostering trust and usability for both moderators and end-users [178].

Privacy protection is shifting from passive anonymization to proactive defense and dynamic monitoring. Techniques such as differential privacy, knowledge distillation, and tag-based prediction reduce raw data exposure, while systems like P-BERT and ProPILE enable real-time detection and blocking of sensitive information. Federated learning enables privacy-preserving model training across distributed systems without sharing raw data, while linguistic steganography provides methods for secure information encoding in collaborative settings [14,170]. Nevertheless, these methods must

 doi:10.20944/preprints202508.1280.v1

balance protection strength, system overhead, and user experience, and their effectiveness against advanced inference and identity linkage attacks remains an open question.

6.2. Implications

Technical Implications. The technical implications of LLM deployment in social media manifest in three interconnected dimensions: content verification capabilities, adversarial content challenges, and platform security vulnerabilities. In terms of content verification, LLMs have transformed how platforms detect and combat false information. For example, Staab et al. [146] and Collier et al. [34] demonstrate how these systems can effectively identify coordinated misinformation campaigns by analyzing subtle patterns across posts and user behaviors. However, this enhanced detection capability comes with its own risks—for example, the same pattern recognition capabilities can be exploited to breach user privacy and reconstruct sensitive information from seemingly innocuous social media interactions [79].

The second dimension concerns the evolving nature of social media content manipulation. LLM-enhanced platforms face increasingly sophisticated forms of synthetic content and automated manipulation. As Duncan et al. [44] suggests, these platforms must balance aggressive content filtering with the risk of false positives that could suppress legitimate discourse. This challenge is particularly apparent in cross-cultural contexts, where content moderation systems must adapt to diverse cultural norms while maintaining consistent integrity standards [93].

The third dimension addresses platform security and scalability. While federated approaches [122] offer promising solutions for privacy-preserving content moderation, they introduce new vulnerabilities in cross-platform coordination. The emergence of sophisticated adversarial techniques targeting social media integrity systems[113] necessitates continuous adaptation of defense mechanisms.

Societal Implications. Based on our review, we also summarize three interconnected dimensions in terms of the societal impact, including platform trust dynamics, information equity, and collective behavior shifts. In the trust dimension, LLMs are fundamentally reshaping how users verify and consume social media content. While these systems enhance automated fact-checking capabilities, they also introduce new challenges to platform credibility. One typical example is that user trust can fluctuate dramatically following high-profile misinformation incidents, particularly when LLM-based detection systems fail to catch sophisticated synthetic content [64]. This trust volatility is further complicated by cross-cultural variations [92]; that is, different societies exhibit varying levels of confidence in automated content verification systems, which further affect the overall effectiveness of platform integrity measures.

The second dimension concerns information equity and accessibility. While LLM-powered content verification promises more democratic access to fact-checking tools, their deployment often creates new forms of information disparities. Specifically, the effectiveness of these systems varies significantly across languages and cultural contexts [44,93], potentially marginalizing users from non-dominant linguistic and cultural backgrounds. For instance, fact-checking systems may perform well on English-language misinformation but fail to detect similar content in Swahili or Tagalog due to limited training data or weaker language modeling support.

The third dimension addresses collective behavior adaptation in response to LLM-enhanced content moderation. Social media users are developing new strategies for information sharing and verification, shaped by their interactions with automated content analysis systems. This has been widely demonstrated in prior studies [77,151], in which researchers have observed how communities adapt their communication patterns to either work with or circumvent LLM-based content filters, and reported the emergence of new forms of collaborative fact-checking that combine human expertise with LLM capabilities.

 doi:10.20944/preprints202508.1280.v1

6.3. Future Research

6.3.1. Information Disorder Detection

The future of information disorder detection and mitigation requires a comprehensive approach integrating technical innovations, adaptive systems, and user-centric considerations.

Technical Foundations and Detection Systems The cornerstone of future information disorder mitigation is the development of robust detection frameworks that fuse multiple data modalities and verification techniques. Multi-modal detection systems integrating text, metadata, and contextual signals show particular promise, demonstrating accuracy improvements of 5-10% over traditional text-only approaches [96]. Similarly, cross-model verification systems, which uses ensemble approaches with multiple LLMs, achieve 8-12% higher detection accuracy for subtle forms of misinformation [45]. These architectures are further enhanced by data synthesis and augmentation techniques, exemplified by FACT-GPT [30] that improve detection capabilities through automated fact-checking and content verification. As adversarial tactics evolve, robust defense mechanisms become crucial. Recent advances in adversarial training and prompt-level hardening, such as Prompt Adversarial Tuning [111], demonstrate promising approaches to enhancing model robustness while maintaining legitimate task performance. These developments can be coupled with improved prompt engineering and detection mechanisms [4,84,91] to address evolving misinformation tactics effectively.

Adaptive Learning and Cultural Context Future systems must emphasize adaptability through continuous learning mechanisms that maintain accuracy on known patterns while quickly adapting to emerging threats. Meta-learning and parameter-efficient tuning approaches have shown potential in reducing annotation costs and accelerating responses to emergent narratives [172]. Integration of human feedback loops, through expert validation and model updates, can also help refine model accuracy and prevent the misclassification of benign content [165]. Cross-cultural and linguistic adaptation represents another critical frontier. Language-agnostic, culturally aware embeddings have shown promise in reducing the gap between high- and low-resource settings [106]. By developing enhanced embeddings and culturally adaptive techniques [55,128], detection systems can achieve effective detection across diverse linguistic and cultural environments.

User Interface and Behavioral Analysis The success of information disorder mitigation depends critically on effective user interaction and behavioral understanding. Future interfaces must clearly communicate detection confidence levels and provide transparent explanations for flagged content. The BiasX framework demonstrates the value of free-text explanations in content moderation, significantly aiding moderators in identifying subtle toxic content [178]. Understanding user behavior patterns, particularly through silent-user and propagation-path modeling [113], enables more effective intervention strategies. Simulation models analyzing propagation dynamics [90,95] further enhance our ability to develop proactive mitigation approaches.

6.3.2. LLM-Enhanced Social Bot Detection

The evolution of LLM-enhanced bot detection necessitates a new generation of detection technologies that are multi-faceted, real-time, and privacy-conscious.

Advanced Detection Architectures The next generation of bot detection systems demands hybrid architectures that leverage multiple detection strategies. Research demonstrates that pipelines merging "digital DNA" temporal signatures with transformer encoders yield superior performance over traditional graph-only detectors [25]. Federated variants of these systems show particular promise in preserving privacy while enabling cross-platform intelligence sharing [168]. The integration of federated learning approaches, as demonstrated by Puppala et al. [121], represents a crucial advancement in enhancing detection capabilities while maintaining robust privacy standards.

Cross-modal and Multilingual Capabilities As bot operations increasingly span multiple languages and media types, future research must prioritize comprehensive cross-lingual and multimodal detection capabilities. Evaluation frameworks such as ETS-MM are advancing joint text-audio-visual embedding analysis [88], while building on foundational work in cross-lingual detection by Panda et

 doi:10.20944/preprints202508.1280.v1

al. [116] and Ahmed et al. [1]. These developments enable more effective understanding and detection of malicious content across linguistic and cultural boundaries. The comprehensive evaluation frameworks pioneered by Gu et al. [58] provide essential benchmarks for assessing multimodal bot detection effectiveness.

Real-time Monitoring and Response The advancement of real-time monitoring capabilities represents a critical direction for future research. Streaming graph transformers have achieved significant breakthroughs in reducing moderation latency to sub-second levels, meeting crucial requirements for live platforms [177]. Recent innovations by Ramamonjisoa et al. [126] and Bonechi et al. [17] in automated moderation systems establish new benchmarks for real-time monitoring and response capabilities. These systems must balance immediate detection with accuracy and resource efficiency.

6.3.3. Privacy Preservation

The future of privacy preservation in LLM-enhanced social media environments demands a multi-layered framework integrating three key components:

Advanced Privacy-Preserving Architectures The foundation of future privacy protection lies in developing sophisticated architectural solutions that can safeguard user data while maintaining system functionality. Differential privacy approaches show particular promise, with recent work by Coffey et al. [33] demonstrating how differentially private stochastic gradient descent during fine-tuning can effectively limit model memorization while preserving utility. Federated learning variants, as explored by Yao et al. [170], offer promising solutions for distributed privacy protection, though they must carefully balance privacy guarantees with system performance. These advances must be complemented by innovative anonymization strategies that can adapt to evolving threats and content sensitivity levels.

Contextual Integrity and Inference Protection A critical challenge in current LLM systems is maintaining contextual privacy across diverse interaction scenarios. The ConfAIde benchmark reveals concerning levels of sensitive context leakage, with up to 57% of tested scenarios showing vulnerability [110]. Future research must focus on developing robust inference-time privacy guards that can dynamically adapt to different contexts while preventing unauthorized information extraction. The CONFAIDE framework [109] provides a foundation for standardizing privacy evaluations, but more sophisticated protection mechanisms are needed to address emerging threats while preserving model utility. Linguistic steganography techniques [14] and advanced PII protection methods [70] show promise in providing granular privacy controls across different linguistic and cultural contexts.

Real-time Monitoring and Compliance As privacy threats evolve and regulatory requirements become more stringent, particularly with frameworks like the EU AI Act [144], real-time privacy monitoring and compliance systems become crucial. Systems like ProPILE [81] demonstrate the potential for proactive privacy protection through continuous monitoring and intervention. Future research must focus on developing explainable systems that can provide clear documentation of privacy measures while adapting to emerging regulatory requirements. This includes creating comprehensive risk assessment frameworks and establishing transparent audit trails that balance privacy protection with accountability.

7. Conclusions

This comprehensive review examines the critical challenges and future directions in social media information integrity within the context of LLM applications. Our systematic analysis reveals three critical challenges that define this landscape: the escalating sophistication of information disorder, the emergence of LLM-enhanced social bots, and the growing complexity of privacy preservation. Through detailed examination of current approaches and their limitations, we have identified how LLMs simultaneously serve as powerful tools for detecting malicious content while potentially enabling more sophisticated forms of synthetic content generation and manipulation. Future developments must focus on several key areas: advancing multi-modal and adaptive detection systems through ensemble approaches; implementing robust privacy-preserving mechanisms to address contextual

 doi:10.20944/preprints202508.1280.v1

privacy leakage in current systems; developing cross-platform coordination capabilities to combat increasingly sophisticated bot activities; and establishing comprehensive regulatory frameworks that balance innovation with ethical considerations. For social media platforms, these challenges necessitate comprehensive strategies combining policy, technical, and user protection measures. Platforms should develop clear guidelines for LLM-generated content, implement robust detection and moderation systems, and establish transparent content labeling mechanisms. Additionally, user education and protection should be prioritized through enhanced verification tools and effective reporting systems. The success of these efforts will depend on the effective integration of technical solutions with policy frameworks, while maintaining transparency and user trust across social media platforms. As LLM technology continues to evolve, the approaches outlined in this review provide a foundation for building more secure, privacy-respecting, and trustworthy social media environments.

References

1. Kawsar Ahmed, Md Osama, Md Sirajul Islam, Md Taosiful Islam, Avishek Das, and Mohammed Moshiul Hoque. 2023. Score_IsAll_you_need at BLP-2023 task 1: A hierarchical classification approach to detect violence inciting text using transformers. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)* (Singapore). Association for Computational Linguistics, Stroudsburg, PA, USA, 185–189.
2. Hani Al-Omari, Malak Abdullah, Ola Al-Titi, and Samira Shaikh. 2019. Justdeep at nlp4if 2019 shared task: propaganda detection using ensemble deep learning models. *EMNLP-IJCNLP* 2019 (2019), 113.
3. Hani Al-Omari, Malak Abdullah, Ola AlTiti, and Samira Shaikh. 2019. JUSTDeep at NLP4IF 2019 task 1: Propaganda detection using ensemble deep learning models. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda* (Hong Kong, China). Association for Computational Linguistics, Stroudsburg, PA, USA.
4. Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo. 2024. Cross-domain fake news detection using a prompt-based approach. *Future Internet* 16, 8 (2024), 286.
5. Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–236.
6. Sawsan Alshattnawi, Amani Shatnawi, Anas M R AlSobeh, and Aws A Magableh. 2024. Beyond word-based model embeddings: Contextualized representations for enhanced social media spam detection. *Appl. Sci. (Basel)* 14, 6 (March 2024), 2254.
7. Sacha Altay, Manon Berriche, Hendrik Heuer, Johan Farkas, and Steven Rathje. 2023. A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review* 4, 4 (2023), 1–34.
8. Dimitris Asimopoulos, Ilias Siniosoglou, Vasileios Argyriou, Thomai Karamitsou, Eleftherios Fountoukidis, Sotirios K Goudos, Ioannis D Moscholios, Konstantinos E Psannis, and Panagiotis Sarigiannidis. 2024. Benchmarking Advanced Text Anonymisation Methods: A Comparative Study on Novel and Traditional Approaches. *arXiv preprint arXiv:2404.14465* (2024).
9. Hadi Askari, Anshuman Chhabra, Bernhard Clemm von Hohenberg, Michael Heseltine, and Magdalena Wojcieszak. 2024. Incentivizing news consumption on social media platforms using large language models and realistic bot accounts. *PNAS Nexus* 3, 9 (Sept. 2024), gae368.
10. Associated Press. 2025. *AP Fact Check*. https://apnews.com/ap-fact-check
11. Maurizio Atzori, Eleonora Calò, Loredana Caruccio, Stefano Cirillo, Giuseppe Polese, and Giandomenico Solimando. 2024. Evaluating password strength based on information spread on social networks: A combined approach relying on data reconstruction and generative models. *Online Soc. Netw. Media* 42, 100278 (Aug. 2024), 100278.
12. Navid Ayoobi, Sadat Shahriar, and Arjun Mukherjee. 2023. The looming threat of fake and llm-generated linkedin profiles: Challenges and opportunities for detection and prevention. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*. 1–10.
13. Jackie Ayoub, X Jessie Yang, and Feng Zhou. 2021. Combat COVID-19 infodemic using explainable natural language processing models. *Information Processing & Management* 58, 4 (2021), 102569.
14. Calvin Bao and Marine Carpuat. 2024. Keep It Private: Unsupervised Privatization of Online Text. *arXiv preprint arXiv:2405.10260* (2024).

 doi:10.20944/preprints202508.1280.v1

15. George-Octavian Bărbulescu and Peter Triantafillou. 2024. To Each (Textual Sequence) Its Own: Improving Memorized-Data Unlearning in Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.
16. Dipto Barman, Ziyi Guo, and Owen Conlan. 2024. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Machine Learning with Applications* (2024), 100545.
17. Simone Bonechi. 2024. Development of an automated moderator for deliberative events. *Electronics (Basel)* 13, 3 (Jan. 2024), 544.
18. Ali Borji and Mehrdad Mohammadian. 2023. Battle of the wordsmiths: Comparing ChatGPT, GPT-4, Claude, and bard. *SSRN Electron. J.* (2023).
19. Robert Brandt. 2023. AI-Assisted Medicine: Possibly Helpful, Possibly Terrifying. *Emergency Medicine News* 45, 7 (2023), 20.
20. Olivia Brown, Robert M Davison, Stephanie Decker, David A Ellis, James Faulconbridge, Julie Gore, Michelle Greenwood, Gazi Islam, Christina Lubinski, Niall G MacKenzie, et al. 2024. Theory-driven perspectives on generative artificial intelligence in business and management. *British Journal of Management* 35, 1 (2024), 3–23.
21. Yunna Cai, Fan Wang, Haowei Wang, and Qianwen Qian. 2023. Public sentiment analysis and topic modeling regarding ChatGPT in mental health on Reddit: Negative sentiments increase over time. *arXiv preprint arXiv:2311.15800* (2023).
22. Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Colin Raffel, et al. 2021. Extracting Training Data from Large Language Models. In *Proceedings of the 30th USENIX Security Symposium*. USENIX Association, 2633–2650.
23. Rosario Catelli, Hamido Fujita, Giuseppe De Pietro, and Massimo Esposito. 2022. Deceptive reviews and sentiment polarity: Effective link by exploiting BERT. *Expert Syst. Appl.* 209, 118290 (Dec. 2022), 118290.
24. Pew Research Center. 2024. *Social Media and News Fact Sheet*. https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/
25. Vaishali Chawla and Yatin Kapoor. 2023. A hybrid framework for bot detection on twitter: Fusing digital DNA with BERT. *Multimed. Tools Appl.* 82, 20 (Aug. 2023), 30831–30854.
26. Ben Chen, Bin Chen, Dehong Gao, Qijin Chen, Chengfu Huo, Xiaonan Meng, Weijun Ren, and Yang Zhou. 2021. Transformer-based language model fine-tuning methods for COVID-19 fake news detection. In *Combating online hostile posts in regional languages during emergency situation: First international workshop, CONSTRAINT 2021, collocated with AAAI 2021, virtual event, February 8, 2021, revised selected papers 1*. Springer, 83–92.
27. Canyu Chen and Kai Shu. 2023. Can LLM-generated misinformation be detected? (2023). 2309.13788 [cs.CL]
28. Tsun-Hin Cheung and Kin-Man Lam. 2023. Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 846–853.
29. Xiaoxiao Chi, Xuyun Zhang, Yan Wang, Lianyong Qi, Amin Beheshti, Xiaolong Xu, Kim-Kwang Raymond Choo, Shuo Wang, and Hongsheng Hu. 2024. Shadow-free membership inference attacks: recommender systems are more vulnerable than you thought. *arXiv preprint arXiv:2405.07018* (2024).
30. Eun Cheol Choi and Emilio Ferrara. 2024. Automated claim matching with large language models: empowering fact-checkers in the fight against misinformation. In *Companion Proceedings of the ACM Web Conference 2024*. 1441–1449.
31. Jeff Christensen, Jared M Hansen, and Paul Wilson. 2024. Understanding the role and impact of Generative Artificial Intelligence (AI) hallucination within consumers' tourism decision-making processes. *Curr. Issues Tourism* (Jan. 2024), 1–16.
32. Cloudflare. 2025. *What is a social media bot?* https://www.cloudflare.com/learning/bots/what-is-a-social-media-bot/ Accessed: March 17, 2025.
33. Sean M Coffey, Joseph W Catudal, and Nathaniel D Bastian. 2024. Differential privacy to mathematically secure fine-tuned large language models for linguistic steganography. In *Assurance and Security for AI-enabled Systems*, Vol. 13054. SPIE, 160–171.
34. Henry Collier. 2024. AI: The Future of Social Engineering! *Proc. Eur. Conf. Inf. Warf. Secur.* 23, 1 (June 2024).
35. Alexis Conneau, Guillaume Lample, Marco Ranzato, Lionel Denoyer, and Hervé Jégou. 2020. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv preprint arXiv:1911.02116* (2020).

 doi:10.20944/preprints202508.1280.v1

36. Nicole A Cooke. 2017. Posttruth, truthiness, and alternative facts: Information behavior and critical information consumption for a new age. *The library quarterly* 87, 3 (2017), 211–221.
37. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. (2019), 4171–4186.
38. Jane Dickson and Thomas S. McCauley. 2023. The Ethics of AI in Mental Health: Safeguarding Patient Privacy in Online Settings. *ACM Transactions on Internet Technology* 23, 2 (2023), 1–24.
39. Dilara Dogan, Bahadir Altun, Muhammed Said Zengin, Mucahid Kutlu, and Tamer Elsayed. 2023. Catch Me If You Can: Deceiving Stance Detection and Geotagging Models to Protect Privacy of Individuals on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 173–184.
40. Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2023. Reducing Privacy Risks in Online Self-Disclosures with Language Models. *arXiv preprint arXiv:2311.09538* (2023).
41. Sunny Duan, Mikail Khona, Abhiram Iyer, Rylan Schaeffer, and Ila Rani Fiete. 2025. Uncovering Latent Memories in Large Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
42. David Dukic, Dominik Keca, and Dominik Stipic. 2020. Are you human? Detecting bots on twitter using BERT. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (sydney, Australia). IEEE.
43. Chris Dulhanty, Jason L Deglint, Ibrahim Ben Daya, and Alexander Wong. 2019. Taking a stance on fake news: Towards automatic disinformation assessment via deep bidirectional transformer language models for stance detection. *arXiv preprint arXiv:1911.11951* (2019).
44. Clay Duncan and Ian Mcculloh. 2023. Unmasking Bias in Chat GPT Responses. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining* (Kusadasi Turkiye). ACM, New York, NY, USA.
45. Mohammed E. Almandouh, Mohammed F Alrahmawy, Mohamed Eisa, Mohamed Elhoseny, and AS Tolba. 2024. Ensemble based high performance deep learning models for fake news detection. *Scientific Reports* 14, 1 (2024), 26591.
46. Mengyi Wang Fangfang Shan, Huifang Sun. 2024. Multimodal Social Media Fake News Detection Based on Similarity Inference and Adversarial Networks. *Computers, Materials & Continua* 79, 1 (2024), 581–605. 1546-2226
47. Shangbin Feng, Zhaoxuan Tan, Herun Wan, Ningnan Wang, Zilong Chen, Binchi Zhang, Qinghua Zheng, Wenqian Zhang, Zhenyu Lei, Shujie Yang, et al. 2022. TwiBot-22: Towards Graph-Based Twitter Bot Detection. In *Advances in Neural Information Processing Systems*.
48. Shangbin Feng, Herun Wan, Ningnan Wang, Zhaoxuan Tan, Minnan Luo, and Yulia Tsvetkov. 2024. What does the bot say? Opportunities and risks of large language models in social media bot detection. (2024). 2402.00371 [cs.CL]
49. Chiarello Filippo, Giordano Vito, Spada Irene, Barandoni Simone, and Fantoni Gualtiero. 2024. Future applications of generative large language models: A data-driven case study on ChatGPT. *Technovation* 133 (2024), 103002.
50. Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 1322–1333.
51. José Antonio García-Díaz and Rafael Valencia-García. 2022. Compilation and evaluation of the Spanish SatiCorpus 2021 for satire identification using linguistic features and transformers. *Complex Intell. Syst.* 8, 2 (April 2022), 1723–1736.
52. Andres Garcia-Silva, Cristian Berrio, and Jose Manuel Gomez-Perez. 2021. Understanding Transformers for Bot Detection in Twitter. (2021). [arxiv]2104.06182
53. Razan Ghanem and Hasan Erbay. 2020. Context-dependent model for spam detection on social networks. *SN Appl. Sci.* 2, 9 (Sept. 2020).
54. Razan Ghanem, Hasan Erbay, and Khaled Bakour. 2023. Contents-based spam detection on social networks using RoBERTa embedding and stacked BLSTM. *SN Comput. Sci.* 4, 4 (May 2023).
55. Masood Ghayoomi. 2023. Enriching contextualized semantic representation with textual information transmission for COVID-19 fake news detection: A study on English and Persian. *Digital Scholarship in the Humanities* 38, 1 (2023), 99–110.
56. Erwin Gielens, Jakub Sowula, and Philip Leifeld. 2025. Goodbye human annotators? Content analysis of social policy debates using ChatGPT. *Journal of Social Policy* (2025), 1–20.

 doi:10.20944/preprints202508.1280.v1

57. Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 US presidential election. *Science* 363, 6425 (2019), 374–378.
58. Tianle Gu, Zeyang Zhou, Kexin Huang, Liang Dandan, Yixu Wang, Haiquan Zhao, Yuanqi Yao, Yujiu Yang, Yan Teng, Yu Qiao, et al. 2024. Mllmguard: A multi-dimensional safety evaluation suite for multimodal large language models. *Advances in Neural Information Processing Systems* 37 (2024), 7256–7295.
59. Qinglang Guo, Haiyong Xie, Yangyang Li, Wen Ma, and Chao Zhang. 2021. Social bots detection via fusing BERT and graph convolutional networks. *Symmetry (Basel)* 14, 1 (Dec. 2021), 30.
60. Ankur Gupta*, School of Information Technology, RGPV, Bhopal, India., Yogendra P S Maravi, Nishchol Mishra, School of Information Technology, RGPV Bhopal, India., and School of Information Technology, RGPV Bhopal, India. 2019. Twitter Spam Detection using Pre-trained Model. *International Journal of Recent Technology and Engineering (IJRTE)* 8, 4 (Nov. 2019), 10620–10623.
61. Fouzi Harrag, Maria Dabbah, Kareem Darwish, and Ahmed Abdelali. 2020. Bert Transformer Model for Detecting Arabic GPT2 Auto-Generated Tweets. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP)*. Association for Computational Linguistics, Barcelona, Spain (Online), 207–214. https://aclanthology.org/2020.wanlp-1.19/
62. Ehtesham Hashmi, Sule Yildirim Yayilgan, Muhammad Mudassar Yamin, Subhan Ali, and Mohamed Abomhara. 2024. Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI. *IEEE Access* 12 (2024), 44462–44480.
63. Katherine Haynes, Hossein Shirazi, and Indrakshi Ray. 2021. Lightweight URL-based phishing detection using natural language processing transformers for mobile devices. *Procedia Comput. Sci.* 191 (2021), 127–134.
64. Dan Heaton, Jeremie Clos, Elena Nichele, and Joel E Fischer. 2024. "The ChatGPT bot is causing panic now – but it'll soon be as mundane a tool as Excel": analysing topics, sentiment and emotions relating to ChatGPT on Twitter. *Pers. Ubiquitous Comput.* (May 2024).
65. Maryam Heidari and James H Jones. 2020. Using BERT to extract topic-independent sentiment features for social media bot detection. In *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (New York, NY, USA). IEEE.
66. Hanjuan Huang, Hsuan-Ting Peng, and Hsing-Kuo Pao. 2023. Fake News Detection via Sentiment Neutralization. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 5780–5789.
67. Jinyuan Huang, Zhao Song, Keliang Li, Shuang Zhang, Sitao Duan, Bo Li, and Haitao Zhao. 2020. InstaHide: Instance-hiding Schemes for Private Discourse on Public Training. In *International Conference on Machine Learning (ICML)*.
68. Kexin Huang, Xiangyang Liu, Qianyu Guo, Tianxiang Sun, Jiawei Sun, Yaru Wang, Zeyang Zhou, Yixu Wang, Yan Teng, Xipeng Qiu, Yingchun Wang, and Dahua Lin. 2024. Flames: Benchmarking Value Alignment of LLMs in Chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, Mexico City, Mexico, 4551–4591.
69. Steve Huntsman, Michael Robinson, and Ludmilla Huntsman. 2024. Prospects for inconsistency detection using large language models and sheaves. *arXiv preprint arXiv:2401.16713* (2024).
70. Sungsoon Jang, Yeseul Cho, Hyeonmin Seong, Taejong Kim, and Hosung Woo. 2024. The Development of a Named Entity Recognizer for Detecting Personal Information Using a Korean Pretrained Language Model. *Applied Sciences* 14, 13 (2024), 5682.
71. Shan Jiang, Miriam Metzger, Andrew Flanagin, and Christo Wilson. 2020. Modeling and measuring expressed (dis) belief in (mis) information. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 315–326.
72. Weiqiang Jin, Ningwei Wang, Tao Tao, Bohang Shi, Haixia Bi, Biao Zhao, Hao Wu, Haibin Duan, and Guang Yang. 2024. A veracity dissemination consistency-based few-shot fake news detection framework by synergizing adversarial and contrastive self-supervised learning. *Scientific Reports* 14, 1 (2024), 19470.
73. Bianca Montes Jones and Marwan Omar. 2023. Detection of twitter spam with language models: A case study on how to use BERT to protect children from spam on twitter. In *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)* (Las Vegas, NV, USA). IEEE, 511–516.
74. Niket Tandon Kandpal, Samuel Bowman, Ethan Perez, and Colin Raffel. 2023. Quantifying Memorization Across Neural Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

 doi:10.20944/preprints202508.1280.v1

75. Debanjana Kar, Mohit Bhardwaj, Suranjana Samanta, and Amar Prakash Azad. 2021. No rumours please! a multi-indic-lingual approach for covid fake-tweet detection. (2021), 1–5.
76. Kornraphop Kawintiranon, Lisa Singh, and Ceren Budak. 2022. Traditional and context-specific spam detection in low resource settings. *Mach. Learn.* 111, 7 (July 2022), 2515–2536.
77. Indra Kertati, Carlos Y T Sanchez, Muhammad Basri, Muhammad Najib Husain, and Hery Winoto Tj. 2023. Public relations' disruption model on chatgpt issue. *J. Studi Komun. (Indones. J. Commun. Stud.)* 7, 1 (March 2023), 034–048.
78. Ashfia Jannat Keya, Md. Anwar Hussen Wadud, M. F. Mridha, Mohammed Alatiyyah, and Md. Abdul Hamid. 2022. AugFake-BERT: Handling Imbalance through Augmentation of Fake News Using BERT to Enhance the Performance of Fake News Classification. *Applied Sciences* 12, 17 (2022). 2076-3417
79. M Mehdi Kholoosi, M Ali Babar, and Roland Croft. 2024. A Qualitative Study on Using ChatGPT for Software Security: Perception vs. Practicality. (2024), 107–117.
80. Myeong Gyu Kim, Minjung Kim, Jae Hyun Kim, and Kyungim Kim. 2022. Fine-tuning BERT models to classify misinformation on garlic and COVID-19 on Twitter. *Int. J. Environ. Res. Public Health* 19, 9 (April 2022), 5126.
81. Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2024. Propile: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems* 36 (2024).
82. Sahas Koka, Anthony Vuong, and Anish Kataria. 2024. Evaluating the Efficacy of Large Language Models in Detecting Fake News: A Comparative Analysis. *arXiv preprint arXiv:2406.06584* (2024).
83. Shubham Kumar, Shivang Garg, Yatharth Vats, and Anil Singh Parihar. 2021. Content based bot detection using bot language model and BERT embeddings. In *2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP)* (Chennai, India). IEEE.
84. Jianqiao Lai, Xinran Yang, Wenyue Luo, Linjiang Zhou, Langchen Li, Yongqi Wang, and Xiaochuan Shi. 2024. RumorLLM: A Rumor Large Language Model-Based Fake-News-Detection Data-Augmentation Approach. *Applied Sciences* 14, 8 (2024). 2076-3417
85. David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.
86. Jaeyoung Lee, Ximing Lu, Jack Hessel, Faeze Brahman, Youngjae Yu, Yonatan Bisk, Yejin Choi, and Saadia Gabriel. 2024. How to Train Your Fact Verifier: Knowledge Transfer with Multimodal Open Models. *arXiv preprint arXiv:2407.00369* (2024).
87. Siyu Li, Jin Yang, and Kui Zhao. 2023. Are you in a masquerade? exploring the behavior and impact of large language model driven social bots in online social networks. *arXiv preprint arXiv:2307.10337* (2023).
88. Wei Li, Jiawen Deng, Jiali You, Yuanyuan He, Yan Zhuang, and Fuji Ren. 2025. ETS-MM: A Multi-Modal Social Bot Detection Model Based on Enhanced Textual Semantic Representation. In *Proceedings of the ACM on Web Conference 2025*. 4160–4170.
89. Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2022. Large Language Models Can Be Strong Differentially Private Learners. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
90. Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2024. Large Language Model Agent for Fake News Detection. *arXiv preprint arXiv:2405.01593* (2024).
91. Yufan Li, Zhan Wang, and Theo Papatheodorou. 2024. Staying vigilant in the Age of AI: From content generation to content authentication. *arXiv preprint arXiv:2407.00922* (2024).
92. Ying Lian, Huiting Tang, Mengting Xiang, and Xuefan Dong. 2024. Public attitudes and sentiments toward ChatGPT in China: A text mining analysis based on social media. *Technol. Soc.* 76, 102442 (March 2024), 102442.
93. Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2024. Investigating bias in llm-based bias detection: Disparities between llms and human perception. *arXiv preprint arXiv:2403.14896* (2024).
94. Xiaohan Liu, Yue Zhan, Hao Jin, Yuan Wang, and Yi Zhang. 2023. Research on the classification methods of social bots. *Electronics (Basel)* 12, 14 (July 2023), 3030.
95. Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. 2024. From Skepticism to Acceptance: Simulating the Attitude Dynamics Toward Fake News. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, Kate Larson (Ed.). International Joint Conferences on Artificial Intelligence Organization, 7886–7894. Human-Centred AI.

 doi:10.20944/preprints202508.1280.v1

96. Yifan Liu, Yaokun Liu, Zelin Li, Ruichen Yao, Yang Zhang, and Dong Wang. 2025. Modality interactive mixture-of-experts for fake news detection. In *Proceedings of the ACM on Web Conference 2025*. 5139–5150.
97. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019). https://arxiv.org/abs/1907.11692
98. Ye Liu, Jiajun Zhu, Kai Zhang, Haoyu Tang, Yanghai Zhang, Xukai Liu, Qi Liu, and Enhong Chen. 2024. Detect, Investigate, Judge and Determine: A Novel LLM-based Framework for Few-shot Fake News Detection. *arXiv preprint arXiv:2407.08952* (2024).
99. Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30. 4765–4774.
100. Hanjia Lyu, Jinfa Huang, Daoan Zhang, Yongsheng Yu, Xinyi Mou, Jinsheng Pan, Zhengyuan Yang, Zhongyu Wei, and Jiebo Luo. 2023. Gpt-4v (ision) as a social media analysis engine. *arXiv preprint arXiv:2311.07547* (2023).
101. SreeJagadeesh Malla and PJA Alphonse. 2022. Fake or real news about COVID-19? Pretrained transformer model to detect potential misleading news. *The European Physical Journal Special Topics* 231, 18 (2022), 3347–3356.
102. J. Martin, P. Johnson, and D. Chang. 2022. AI and Mental Health: Privacy Challenges and Solutions. *IEEE Transactions on Computational Social Systems* 9, 3 (2022), 180–192.
103. Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. The limits of word level differential privacy. *arXiv preprint arXiv:2205.02130* (2022).
104. Nikhil Mehta and Dan Goldwasser. 2024. Using RL to identify divisive perspectives improves LLMs abilities to identify communities on social media. *arXiv preprint arXiv:2406.00969* (2024).
105. Meta Platforms, Inc. 2023. Facebook Privacy Policy. Available at https://www.facebook.com/privacy/policy, accessed March 2024.
106. Zhongtao Miao, Qiyu Wu, Kaiyan Zhao, Zilong Wu, and Yoshimasa Tsuruoka. 2024. Enhancing cross-lingual sentence embedding for low-resource languages with word alignment. *arXiv preprint arXiv:2404.02490* (2024).
107. Maria Milkova, Maksim Rudnev, and Lidia Okolskaya. 2023. Detecting value-expressive text posts in Russian social media. *arXiv preprint arXiv:2312.08968* (2023).
108. Helen Milner and Michael Baron. 2023. Establishing an optimal online phishing detection method: Evaluating topological NLP transformers on text message data. *Journal of Data Science and Intelligent Systems* 2, 1 (July 2023), 37–45.
109. Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. Quantifying privacy risks of masked language models using membership inference attacks. *arXiv preprint arXiv:2203.03929* (2022).
110. Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884* (2023).
111. Yichuan Mo, Yuji Wang, Zeming Wei, and Yisen Wang. 2024. Fight back against jailbreaking via prompt adversarial tuning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
112. Robert T. Morris, Florian Tramer, and Nicholas Carlini. 2024. Language Model Inversion: Recovering Training Data from Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
113. Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 1732–1742.
114. Helen Nissenbaum. 2019. Contextual Integrity Up and Down the Data Food Chain. *Theoretical Inquiries in Law* 20, 1 (2019), 221–256.
115. Nikolaos Panagiotou, Antonia Saravanou, and Dimitrios Gunopulos. 2021. News Monitor: A framework for exploring news in real-time. *Data (Basel)* 7, 1 (Dec. 2021), 3.
116. Subhadarshi Panda and Sarah Ita Levitan. 2021. Detecting multilingual COVID-19 misinformation on social media via contextualized embeddings. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda* (Online). Association for Computational Linguistics, Stroudsburg, PA, USA.

 doi:10.20944/preprints202508.1280.v1

117. Constantinos Patsakis and Nikolaos Lykousas. 2023. Man vs the machine in the struggle for effective text anonymisation in the age of large language models. *Scientific Reports* 13, 1 (2023), 16026.
118. Protik Bose Pranto, Syed Zami-Ul-Haque Navid, Protik Dey, Gias Uddin, and Anindya Iqbal. 2022. Are you misinformed? a study of covid-related fake news in bengali on facebook. *arXiv preprint arXiv:2203.11669* (2022).
119. Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833* (2022).
120. Haritz Puerto, Martin Gubri, Sangdoo Yun, and Seong Joon Oh. 2024. Scaling Up Membership Inference: When and How Attacks Succeed on Large Language Models. *arXiv preprint arXiv:2411.00154* (2024).
121. Sai Puppala, Ismail Hossain, Md Jahangir Alam, and Sajedul Talukder. 2024. FLASH: Federated Learning-Based LLMs for Advanced Query Processing in Social Networks through RAG. (2024), 281–293.
122. Sai Puppala, Ismail Hossain, Md Jahangir Alam, and Sajedul Talukder. 2024. SocFedGPT: Federated GPT-based Adaptive Content Filtering System Leveraging User Interactions in Social Networks. *arXiv preprint arXiv:2408.05243* (2024).
123. Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical Multi-modal Contextual Attention Network for Fake News Detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 153–162. 9781450380379
124. Kristina Radivojevic, Nicholas Clark, and Paul Brenner. 2024. LLMs Among Us: Generative AI participating in digital discourse. *Proceedings of the AAAI Symposium Series* 3, 1 (May 2024), 209–218.
125. Sarika S Raga and Chaitra B. 2022. A bert model for sms and twitter spam ham classification and comparative study of machine learning and deep learning technique. In *2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE)* (MANGALORE, India). IEEE.
126. David Ramamonjisoa and Shuma Suzuki. 2024. Comments Analysis in Social Media based on LLM Agents. In *CS & IT Conference Proceedings*, Vol. 14.
127. Junaid Rashid, Jungeun Kim, and Anum Masood. 2024. Unraveling the Tangle of Disinformation: A Multimodal Approach for Fake News Identification on Social Media. In *Companion Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 1849–1853. 9798400701726
128. V Rathinapriya and J Kalaivani. 2024. Adaptive weighted feature fusion for multiscale atrous convolution-based 1DCNN with dilated LSTM-aided fake news detection using regional language text information. *Expert Systems* (2024), e13665.
129. Prismahardi Aji Riyantoko, Tresna Maulana Fahrudin, Dwi Arman Prasetya, Trimono Trimono, and Tahta Dari Timur. 2022. Analisis Sentimen Sederhana Menggunakan Algoritma LSTM dan BERT untuk Klasifikasi Data Spam dan Non-Spam. *PROSIDING SEMINAR NASIONAL SAINS DATA* 2, 1 (Dec. 2022), 103–111.
130. Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023. Benchmarking the Generation of Fact Checking Explanations. *Transactions of the Association for Computational Linguistics* 11 (2023), 1250–1264.
131. Marko Sahan, Vaclav Smidl, and Radek Marik. 2021. Active Learning for Text Classification and Fake News Detection. In *2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*. 87–94.
132. Siva Sai. 2020. Siva at WNUT-2020 task 2: Fine-tuning transformer neural networks for identification of informative covid-19 tweets. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)* (Online). Association for Computational Linguistics, Stroudsburg, PA, USA.
133. Amine Sallah, El Arbi Abdellaoui Alaoui, Said Agoujil, Mudasir Ahmad Wani, Mohamed Hammad, Yassine Maleh, and Ahmed A Abd El-Latif. 2024. Fine-tuned understanding: Enhancing social bot detection with transformer-based classification. *IEEE Access* 12 (2024), 118250–118269.
134. Kanti Singh Sangher, Archana Singh, and Hari Mohan Pandey. 2024. LSTM and BERT based transformers models for cyber threat intelligence for intent identification of social media platforms exploitation from darknet forums. *Int. J. Inf. Technol.* 16, 8 (Dec. 2024), 5277–5292.
135. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).
136. Ali Satvaty, Suzan Verberne, and Fatih Turkmen. 2024. Undesirable Memorization in Large Language Models: A Survey. In *arXiv preprint arXiv:2410.02650*.
137. Isabel Segura-Bedmar and Santiago Alonso-Bartolome. 2022. Multimodal fake news detection. *Information* 13, 6 (2022), 284.

 doi:10.20944/preprints202508.1280.v1

138. Samaneh Shafee, Alysson Bessani, and Pedro M Ferreira. 2025. Evaluation of LLM-based chatbots for OSINT-based Cyber Threat Awareness. *Expert Systems with Applications* 261 (2025), 125509.
139. Siddhant Bikram Shah, Surendrabikram Thapa, Ashish Acharya, Kritesh Rauniyar, Sweta Poudel, Sandesh Jain, Anum Masood, and Usman Naseem. 2024. Navigating the web of disinformation and misinformation: Large language models as double-edged swords. *IEEE Access* (2024), 1–1.
140. Filipo Sharevski, Jennifer Vander Loop, Peter Jachim, Amy Devine, and Emma Pieroni. 2023. Talking abortion (mis) information with chatgpt on tiktok. In *2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 594–608.
141. Utkarsh Sharma, Prateek Pandey, and Shishir Kumar. 2022. A transformer-based model for evaluation of information relevance in online social-media: A case study of Covid-19 media posts. *New Gener. Comput.* 40, 4 (Jan. 2022), 1029–1052.
142. Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Mining disinformation and fake news: Concepts, methods, and recent advancements. *Disinformation, misinformation, and fake news in social media: Emerging research challenges and opportunities* (2020), 1–19.
143. Utsav Shukla, Manan Vyas, and Shailendra Tiwari. 2023. Raphael at ArAIEval shared task: Understanding persuasive language and tone, an LLM approach. In *Proceedings of ArabicNLP 2023* (Singapore (Hybrid)). Association for Computational Linguistics, Stroudsburg, PA, USA, 589–593.
144. Nathalie A Smuha. 2025. Regulation 2024/1689 of the Eur. Parl. & Council of June 13, 2024 (EU Artificial Intelligence Act). *International Legal Materials* (2025), 1–148.
145. Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. AI model GPT-3 (dis) informs us better than humans. *Science Advances* 9, 26 (2023), eadh1850.
146. Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298* (2023).
147. Andrei Stipiuc. 2024. Romanian Media Landscape in 7 Journalists' Facebook Posts: A ChatGPT Sentiment Analysis. *SAECULUM* 57, 1 (2024), 20–46.
148. Ningxin Su, Chenghao Hu, Baochun Li, and Bo Li. 2024. TITANIC: Towards Production Federated Learning with Large Language Models. In *Proceedings of IEEE INFOCOM*.
149. B N[1] Supriya and CB Akki. 2022. P-BERT: Polished Up Bidirectional Encoder Representations from Transformers for Predicting Malicious URL to Preserve Privacy. In *2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. IEEE, 1–6.
150. Yingjie Tian and Yuhao Xie. 2024. Artificial cheerleading in IEO: Marketing campaign or pump and dump scheme. *Inf. Process. Manag.* 61, 1 (Jan. 2024), 103537.
151. Ahmed Tlili, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T Hickey, Ronghuai Huang, and Brighter Agyemang. 2023. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learn. Environ.* 10, 1 (Feb. 2023).
152. Christopher K Tokita, Kevin Aslett, William P Godel, Zeve Sanderson, Joshua A Tucker, Jonathan Nagler, Nathaniel Persily, and Richard Bonneau. 2024. Measuring receptivity to misinformation at scale on a social media platform. *PNAS nexus* 3, 10 (2024), pgae396.
153. Ben Treves, Md Rayhanul Masud, and Michalis Faloutsos. 2023. RURLMAN: Matching Forum Users Across Platforms Using Their Posted URLs. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*. 484–491.
154. Milena Tsvetkova, Taha Yasseri, Niccolo Pescetelli, and Tobias Werner. 2024. A new sociology of humans and machines. *Nature Human Behaviour* 8, 10 (Oct. 2024), 1864–1876. 2397-3374 http://dx.doi.org/10.1038/s41562-024-02001-8
155. UNHCR. 2021. *Using Social Media in Community-Based Protection*. Retrieved January, 2021 from https://www.unhcr.org/innovation/wp-content/uploads/2021/01/Using-Social-Media-in-CBP.pdf
156. Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science* 359, 6380 (2018), 1146–1151.
157. Nazmi Ekin Vural and Sefer Kalaman. 2024. Using Artificial Intelligence Systems in News Verification: An Application on X. *İletişim Kuram ve Araştırma Dergisi* 67 (2024), 127–141.
158. Xiao Wang, Jinyuan Sun, Jinyuan Zhang, Neil Shah, and Bo Li. 2024. Prompt Inversion: Leveraging Attention-Based Inversion to Recover Prompts from Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
159. Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework for research and policymaking*. Vol. 27. Council of Europe Strasbourg.

 doi:10.20944/preprints202508.1280.v1

160. Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2018. Learning Neural Templates for Text Generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 3174–3187.
161. Yunfei Xing, Justin Zuopeng Zhang, Guangqing Teng, and Xiaotang Zhou. 2024. Voices in the digital storm: Unraveling online polarization with ChatGPT. *Technology in Society* 77 (2024), 102534. 0160-791X
162. Guangxia Xu, Daiqi Zhou, and Jun Liu. 2021. Social network spam detection based on ALBERT and combination of Bi-LSTM with self-attention. *Secur. Commun. Netw.* 2021 (April 2021), 1–11.
163. Junhao Xu, Longdi Xian, Zening Liu, Mingliang Chen, Qiuyang Yin, and Fenghua Song. 2024. The future of combating rumors? retrieval, discrimination, and generation. *arXiv preprint arXiv:2403.20204* (2024).
164. Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R Fung, and Heng Ji. 2024. LEMMA: towards LVLM-enhanced multimodal misinformation detection with external knowledge augmentation. *arXiv preprint arXiv:2402.11943* (2024).
165. Hongwei Yan, Liyuan Wang, Kaisheng Ma, and Yi Zhong. 2024. Orchestrate latent expertise: Advancing online continual learning with multi-level supervision and reverse self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23670–23680.
166. Chang Yang, Peng Zhang, Wenbo Qiao, Hui Gao, and Jiaming Zhao. 2023. Rumor detection on social media with crowd intelligence and ChatGPT-assisted networks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 5705–5717.
167. Kaicheng Yang and Filippo Menczer. 2024. Anatomy of an AI-powered malicious social botnet. *J. Quant. Descr. Digit. Media* 4 (May 2024).
168. Yingguang Yang, Renyu Yang, Hao Peng, Yangyang Li, Tong Li, Yong Liao, and Pengyuan Zhou. 2023. FedACK: Federated adversarial contrastive knowledge distillation for cross-lingual and cross-model social bot detection. In *Proceedings of the ACM Web Conference 2023*. 1314–1323.
169. Xin Yao, Tianchi Huang, Chenglei Wu, Rui-Xiao Zhang, and Lifeng Sun. 2019. Federated learning with additional mechanisms on clients to reduce communication costs. *arXiv preprint arXiv:1908.05891* (2019).
170. Yuhang Yao, Jianyi Zhang, Junda Wu, Chengkai Huang, Yu Xia, Tong Yu, Ruiyi Zhang, Sungchul Kim, Ryan Rossi, Ang Li, et al. 2024. Federated large language models: Current progress and future directions. *arXiv preprint arXiv:2409.15723* (2024).
171. Junshuai Yu, Qi Huang, Xiaofei Zhou, and Ying Sha. 2020. IARNet: An Information Aggregating and Reasoning Network over Heterogeneous Graph for Fake News Detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*. 1–9.
172. Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. MetaAdapt: Domain adaptive few-shot misinformation detection via meta learning. *arXiv preprint arXiv:2305.12692* (2023).
173. Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin Vechev. 2024. A Synthetic Dataset for Personal Attribute Inference. *arXiv preprint arXiv:2406.07217* (2024).
174. Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. 2023. Low-cost high-power membership inference attacks. *arXiv preprint arXiv:2312.03262* (2023).
175. Jinyuan Zhang, Xinyue Chen, Xuechen Li, and Cho-Jui Hsieh. 2024. Membership Inference Attacks against Fine-tuned Large Language Models via Self-prompt Calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*.
176. Tong Zhang, Di Wang, Huanhuan Chen, Zhiwei Zeng, Wei Guo, Chunyan Miao, and Lizhen Cui. 2020. BDANN: BERT-Based Domain Adaptation Neural Network for Multi-Modal Fake News Detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*. 1–8.
177. Xiang Zhang, Yufei Cui, Chenchen Fu, Zihao Wang, Yuyang Sun, Xue Liu, and Weiwei Wu. 2025. Transtreaming: Adaptive Delay-aware Transformer for Real-time Streaming Perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 10185–10193.
178. Yiming Zhang, Sravani Nanduri, Liwei Jiang, Tongshuang Wu, and Maarten Sap. 2023. Biasx:" thinking slow" in toxic content moderation with explanations of implied social biases. *arXiv preprint arXiv:2305.13589* (2023).
179. Yizhou Zhang, Karishma Sharma, Lun Du, and Yan Liu. 2024. Toward mitigating misinformation and social media manipulation in LLM era. In *Companion Proceedings of the ACM on Web Conference 2024* (Singapore Singapore), Vol. 19. ACM, New York, NY, USA, 1302–1305.
180. Chenye Zhao and Cornelia Caragea. 2021. Knowledge distillation with BERT for image tag-based privacy prediction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. 1616–1625.

 doi:10.20944/preprints202508.1280.v1

181. Xinyi Zhou, Ashish Sharma, Amy X Zhang, and Tim Althoff. 2024. Correcting misinformation on social media with a large language model. *arXiv preprint arXiv:2403.11169* (2024).
182. Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–40.