Factuality Challenges in the Era of Large Language Models

Isabelle Augenstein[1], Timothy Baldwin[2], Meeyoung Cha[3], Tanmoy Chakraborty[*4], Giovanni Luca Ciampaglia[5], David Corney[6], Renee DiResta[7], Emilio Ferrara[8], Scott Hale[9], Alon Halevy[10], Eduard Hovy[11], Heng Ji[12], Filippo Menczer[13], Ruben Miguez[14], Preslav Nakov[2], Dietram Scheufele[15], Shivam Sharma[4], and Giovanni Zagni[16]

[1]University of Copenhagen, Nørregade 10, 1172 København, Denmark
[2]Mohamed bin Zayed University of Artificial Intelligence, Masdar City, Abu Dhabi, 7909, United Arab Emirates
[3]Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon, South Korea
[4]Indian Institute of Technology Delhi, New Delhi, 110016, India
[5]University of Maryland, College Park, Maryland 20742, USA
[6]Full Fact, 17 Oval Way, London, SE11 5RR, United Kingdom
[7]Stanford University, 450 Jane Stanford Way, Stanford, CA 94305, USA
[8]University of Southern California, Los Angeles, CA 90007, USA
[9]University of Oxford, Broad St, Oxford OX1 3AZ, United Kingdom
[10]Meta AI, 1 Hacker Way, Menlo Park, CA 94025, USA
[11]Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA
[12]University of Illinois Urbana-Champaign, 506 S. Wright St. Urbana, IL 61801-3633, USA
[13]Indiana University, 1015 E 11th St., Bloomington, IN 47408, USA
[14]Newtrales, C/Vandergoten 1, 28014 Madrid, Spain
[15]University of Wisconsin, Madison, WI, USA
[16]Pagella Politica/Facta, viale Monza 259/265, Milano, 20125, Italy

ABSTRACT

The emergence of tools based on Large Language Models (LLMs), such as OpenAI's ChatGPT, Microsoft's Bing Chat, and Google's Bard, has garnered immense public attention. These incredibly useful, natural-sounding tools mark significant advances in natural language generation, yet they exhibit a propensity to generate false, erroneous, or misleading content — commonly referred to as *hallucinations*. Moreover, LLMs can be exploited for malicious applications, such as generating false but credible-sounding content and profiles at scale. This poses a significant challenge to society in terms of the potential deception of users and the increasing dissemination of inaccurate information. In light of these risks, we explore the kinds of technological innovations, regulatory reforms, and AI literacy initiatives needed from fact-checkers, news organizations, and the broader research and policy communities. By identifying the risks, the imminent threats, and some viable solutions, we seek to shed light on navigating various aspects of veracity in the era of generative AI.

1 Introduction

In his pioneering work on information theory published in 1948[1], Claude Shannon proposed simple statistical language models based on letter and word frequencies capable of generating text that, although devoid of meaning, resembled human language. With the vast amounts of digital text available for training and thanks to advances in computation, large language models (LLMs) today are capable of memorizing and reasoning about natural language[2]; they can generate text that human readers find increasingly hard to distinguish from actual human text. From early LLMs such as GPT[3] to more recent models such as GPT-4[4] and LLaMA 2[5], several iterations of these technologies have achieved unprecedented sophistication in natural language understanding and generation ability. Despite the differences in their architectures, training data, and algorithms, LLMs principally aim to address the same task, known as "next-word prediction" — given a sequence of words as a prompt, what word(s) should follow?

LLMs were brought to the general public's attention in late 2022 with the release of OpenAI's ChatGPT[6], a chatbot based on LLM technology specifically trained to generate responses as part of a conversation. Analyst reports found that ChatGPT reached 100 million monthly active users only two months after its launch, making it the fastest-growing consumer application

[*] Corresponding author, Email: tanchak@iitd.ac.in

in history[7]. Subsequent innovations by large tech companies such as Microsoft, Meta, and Google[8] pushed the technology's profile even further. These were then followed by Alpaca[9] and Vicuna[10], open-source alternatives created by instruction-tuning Meta's LLaMA base model[5] with conversations from ChatGPT. Similarly, Claude[11], Falcon[12], Jurassic[13], and Jais[14] were developed through instruction-tuning of their own individual LLM models.

The proliferation of LLMs has led to concerns about an arms race in Generative AI (GenAI) among leading tech companies. In 2020, prior to eventually open-sourcing the second release of their generative pre-trained transformer (GPT) series, GPT-2[15], OpenAI cited the potential for societal harm as one of the reasons for not releasing the source code of their first model, GPT[16]. Since then, the industry has largely abandoned such a cautious stance. As a result, LLMs are rapidly gaining popularity among a vast number of professionals and everyday people with limited or no technical knowledge. On the research front, a recent survey indicated a significant uptick in the number of papers containing the term "large language model" in their title or abstract[17].

The year 2022 brought a dramatic increase in the parameter size of these models, reminiscent of Moore's law, and a qualitative leap in commonsense reasoning[17] and knowledge generalization[18]. Some studies also point out that LLMs with billion-scale parameters have *emergent abilities* that were not present in smaller models[19]. Training these large models requires extensive GPU resources. As a result, a handful of well-funded tech companies and organizations dominate technological advances. This makes LLMs difficult to study from outside. Nonetheless, testing and validating technology is critical to ensuring its reliability, safety, and effectiveness, especially for fast-adopted and widely-used technologies like LLMs.

We focus on one of the potential risks of LLMs in generating inaccurate, misleading, false, or entirely fabricated content. Even though LLMs are developed with good intentions and applied to perform useful tasks, they can generate factually false statements, known as *hallucinations*[20] and *hallucinatory explanations* in defense of such statements. Although widely used, the term "hallucination" may be misleading and encourages anthropomorphism[21]. Thus, discerning fact from fiction in LLM-generated content is difficult, and it is further complicated by the use of eloquent language and confident tone[22].

The implications of LLM hallucinations and the resulting unsolicited justifications are especially concerning in critical and sensitive fields. For example, chatbots may become particularly appealing in information-seeking and question-answering tasks related to public health, given that it is easier for many people to receive health advice from a chatbot than from a physician[23]. Yet, LLMs typically lack access to updated and reliable information sources. A study involving 6,594 real-world chatbot interactions during the COVID-19 pandemic revealed that 30% of users queried using the keyword "COVID-19"[24]. This is concerning, given the unsuitability of present chatbot systems in addressing emerging topics such as COVID-19, both before the pandemic[25] and in hindsight[26]. A recent expert evaluation revealed that ChatGPT ended up giving a "definitive answer to something that is not totally agreed on" on the subject of dialysis for lithium poisoning[26]. Although advanced agents such as ChatGPT, given their release timing, could potentially provide solid answers to basic health questions with well-established scientific consensus[27], they are ill-equipped to handle evolving scientific consensus on controversial or emerging topics.

An even larger concern arises when LLMs are used with malicious intent. There have been several discussions about applications of LLMs that put individuals at risk, such as the generation of inauthentic or misleading content at scale[28]. For example, when ChatGPT was prompted to improve the language of a poorly written phishing email, it not only corrected it but also, without any direct user request, added additional content at the end, asking for money[29]. These capabilities will likely extend beyond mere "proof-of-concept" demonstrations, underscoring the risk of propagating false or manipulative information, deliberately or accidentally, with negative consequences for politics, finance, health, and society at large.

LLM-based chatbots also pose challenges to fact-checking. Fact-checkers now need to vet a surge of persuasive yet unreliable text, especially false news generated with the help of LLMs[30] and spread through social media, including by LLM-based social bots[31]. There are also questions about whether and how the public can over-rely on chatbots such as ChatGPT as research tools for verifying information[32,33]. Search engines such as Google and Bing have historically been seen as reliable gateways to authoritative information sources; while in certain cases, search engines have helped amplify disinformation[34], they are more reliable as they give a clear indication of their sources, unlike ChatGPT. However, as LLM-based chatbots become increasingly used for information-seeking, there is a risk that the public may receive unreliable information through a modality that has traditionally been trustworthy[32]. Recent incidents have highlighted the dangers of this trend, such as Google Bard falsely claiming the James Webb telescope's discovery of an exoplanet image[35] and Replika disseminating misinformation by suggesting that Bill Gates was the architect of COVID-19[36]. These occurrences underscore the delicate balance between the risks and benefits of such advanced technologies.

In summary, LLMs have the potential to mislead the public in novel ways, often in combination with or within environments such as search engines, in which users have come to expect (and assume) accuracy. Moreover, the datasets used to train these models can introduce biases, magnifying specific viewpoints while suppressing others[37], thus raising concerns about their widespread adoption. This article explores how LLMs like ChatGPT affect fact-checking and truthfulness, emphasizing the tangible risks associated with their hallucinatory capabilities and prospects for malicious use. It also discusses potential opportunities and strategies for addressing these challenges and the limitations of current solutions, concluding with key objectives for individuals, organizations, and governments, advocating the harnessing of LLM benefits while reducing the risks.

2 Risks, Threats, and Challenges

There are two sides to the information veracity problem of LLMs. On the one hand, LLMs developed with good intentions may nevertheless be plagued by unreliable information. This could be due to unreliable, inconsistent, incomplete, or missing training data or due to hallucinations. As a result, chatbots or AI-enhanced search may give inaccurate responses on sensitive topics such as finance or health, as they may generate false claims or explanations. This is a risk because users may act upon, share, and/or quote these responses later.

On the other hand, AI chatbots may be intentionally used for malicious services or activities, such as scam emails, disinformation websites, or bot feeds. AI-generated fabricated news websites, for example, are known to have tens of thousands of followers on social media, directly reaching social media users worldwide. The two sides are not independent of each other; disinformation from low-credibility sources supported by malicious GenAI applications is made available as training data for LLMs[38]. While the focus of this article is on textual output, it is worth noting that AI tools with visual capabilities, such as Dall-E[39], MidJourney[40] and Stable Diffusion[41] also pose substantial challenges by enabling the creation of deepfakes and false images at an unprecedented scale[42]. Malicious actors can combine fake visuals with AI-generated text to support false claims or to create fake social media profiles on a large scale.

Traditional fact-checking has evolved to ensure the reliability of information[43], but GenAI poses real threats. The hidden, private nature of chatbot conversations and the potential for misinformation underscore the importance of AI literacy and user awareness. Below, we briefly outline multiple risk factors, imminent threats, and the associated challenges that arise from direct interaction with chatbots and their deliberate malicious use.

2.1 Factuality Challenges

The ubiquitous adoption of AI chatbots can aggravate one of our most pressing challenges — the quality and reliability of information in the digital age. Chatbots can generate false or misleading content that can be quickly disseminated and amplified through various platforms and channels. Some of the main limitations and risks of chatbots are outlined below.

Undersourcing: LLM-generated content tends to be coherent but lacks credible sourcing[26]. This issue extends beyond chatbots to generative search engines. Studies scrutinizing generative search engines also found significant problems with proper citations, with nearly half of the generated sentences lacking citations and only three-quarters of those that have them actually supporting their claims[44]. These findings raise concerns, considering the growing trust of consumers in their reliability.

Truthfulness: Despite their significant strides in natural language generation, LLMs tend to generate undesirable text hallucinations that include nonsensical or significantly divergent output, incoherent content, and factual inaccuracies[45]. Although LLMs excel in creative writing and basic explanations, they struggle with factual accuracy. This has led to their temporary ban on platforms such as Stack Overflow[46]. Multiple studies highlighted ChatGPT's overall lack of trustworthiness[26]. In clinical contexts, it demonstrated higher accuracy (80%) in clinical questions compared to evidence-based ones (36%)[47]. In another study, ChatGPT was shown to perform better in basic clinical science tests but less effectively in specialized domains, while GPT-4 showed greater consistency between various types of tests[48]. In summary, ChatGPT is not a reliable substitute for actual physicians. Overall, LLMs are better at deductive reasoning and less so at inductive reasoning[20], which could affect their reliability for fact-based reasoning tasks[49].

Confident Tone: The text generated by chatbots exudes confidence, which often makes them appear as "authoritative liars." Their persuasive narratives can deceive readers into believing false or meaningless information due to the absence of uncertainty or hedging expressions. Language models tailored for conversation tend to maintain a confident tone even when their accuracy is compromised. Addressing this challenge is difficult because LLMs lack measures of inherent factuality and ways to express related uncertainty. While chatbots such as ChatGPT attempt to emulate uncertainty, most safety checks rely on hard-coded rules based on learned prompts, which can be easily circumvented.

Fluent Style: Chatbots communicate from the first person's perspective and even digress at times like humans being, which encourages anthropomorphism. In addition, the coherent and fluent writing style of chatbots can be persuasive to humans, even on controversial political issues[50]. Numerous studies have shown that fluency shapes the perception of truth "among intelligent people, despite contradictory knowledge, for claims from unreliable sources"[51]. Further studies are needed to examine how users perceive chatbots and their capabilities in terms of AI literacy.

Direct Use: In traditional settings, misinformation claims gain prominence and eventually catch the attention of fact-checkers who can debunk them. However, chatbot interactions are proprietary and occur privately and without mediation, which makes it harder to discern their validity. The misinformation a chatbot conveys may be buried among many other true and false claims, adding to the difficulty of detection and correction. WhatsApp has partnered with a number of fact-checking organisations to create tiplines[52], where users can manually request verification of suspicious claims. Still, this manual process risks being overwhelmed by GenAI content.

Ease of Access: Competition in LLMs is driving fast technological advances. Meta's LLaMA, which came out as a competitor for ChatGPT, is now accessible for personalized use on standard laptop hardware[53] unlike ChatGPT, which is restricted via a rate-limited API. Recent open-sourcing of models such as LLAMA 2[5], Falcon[54], and Jais[14] offer freely downloadable alternatives to encourage transparent and democratized technological development. The ease of access to LLMs can accelerate innovation, but it poses significant risks and challenges when used with malicious intent, as discussed in Section 2.2. Tracing such malicious use will be more challenging with the growing diversity of available models.

Halo Effect: A model's proficiency in addressing one topic might lead users to believe it can excel in any open-domain conversations, an example of a cognitive bias known as the *halo effect*.[55]. Such an assumption is risky for users who seek information on new and urgent topics. For example, a chatbot trained before 2019 may not know anything about COVID-19 and can give wrong responses during a health crisis. Even a large, up-to-date training dataset only reflects a tiny and biased selection of human knowledge.

Public Perception: An LLM may seem like a reliable "knowledge base" to the general public. This can lead some users to consider LLMs as enhanced versions of search engines capable of delivering the best answers. However, this is risky, as demonstrated by a recent incident in which an attorney presented to the court fake cases generated by ChatGPT[56]. Responses presented in the form of answers, rather than a collection of diverse links offering varying perspectives, may produce either inaccurate or biased information. Therefore, users must learn how LLMs operate and not trust responses blindly.

Unreliable Evaluation: Another pressing issue is the unreliable evaluation of LLMs[57]. Assessing subjective factors such as "factuality" and "truthfulness" is a complex task beyond predefined and annotated benchmarks such as BIG-bench[58], GLUE[59], and SuperGLUE[60]. Specialized datasets have been developed to measure the factuality of LLMs, such as TruthfulQA[61], which contains expert-crafted questions that quantify model misconception on the topics of health, law, finance, and politics. There have also been some specialized evaluation measures such as FactScore, which evaluate factuality with respect to Wikipedia (e.g., biography pages)[62]. Benchmark datasets alone are insufficient, as LLMs may have been polluted by encountering the same or similar data during their training[63]. Newer metrics such as GPTScore[64], G-Eval[65], and SelfCheckGPT[66] attempt to address this shortcoming in evaluations, albeit with limitations due to aspects such as positional and numeric biases, stochastic inferencing, and self-preferencing[67]. As of now, LLMs have demonstrated mixed performance on misinformation detection tasks[20, 33], involving test sets consisting of scientific and social claims related to COVID-19[68]. Maintaining accurate ground-truth references, particularly for factuality assessment, is costly, subject to data drift, and remains an open research problem.

2.2 Threats Posed by Malicious LLM Usage

The widespread availability of LLMs empowers people with ideas to craft compelling and elegantly written content and persuasive arguments, a skill traditionally held by experienced writers. Even though all technologies have both good and bad applications, it is important to foresee potential harmful applications of LLMs, as outlined below.

Personalized Attacks: LLMs can generate text that aligns with the context of the ongoing conversation. This capability can be exploited for malicious purposes, where a user's prior statements, such as those from emails or social media, can be incorporated into a prompt to generate disinformation, phishing messages, harassment or other harmful content on a large scale. The content can be tailored to appear credible, personalized, and targeted at specific individuals or groups and can be mixed with factual statements to make it more persuasive. The possibility of data leakage increases these risks. Cyberhaven, a firm that provides data tracing and security solutions, recently analyzed the engagements of their clients' employees with ChatGPT[69] and found that 10.8% had used ChatGPT at work, 8.6% had shared company data via prompts, and 4.7% had accidentally disclosed confidential information. The risk of data leakage is evidenced by the inadvertent disclosure of payment details from 1.2% of ChatGPT Plus subscribers. Open-source LLMs can be manipulated to extract such private information, which can then be used for more effective phishing attacks and other scams.

Style Impersonation: Fine-tuned LLMs will be able to generate text that emulates the style of any person, providing access to relevant training data. Consequently, it will be relatively simple to train a model to generate text that mimics the writing style of specific individuals, such as journalists, fact-checkers, politicians, regulators, and more. This content could then be distributed on social media platforms in an effort to undermine the credibility of perceived adversaries.

Bypassing Detection: Fact-checkers prioritize monitoring and verifying widely circulated claims. For example, a piece of misinformation that has circulated a thousand times is scrutinized more than once by only ten people. However, GenAI tools can generate infinite variations of the same content. Even if each variant reaches a small number of people, the cumulative impact could add to that of a highly viral piece while remaining invisible to fact-checkers. This automated diversification could undermine the safety and integrity efforts of many social media platforms, such as Meta's third-party fact-checking program[70].

Fake Profiles: The ability to generate credible-sounding fake profiles and content at scale will impact social media influencing, making social bots a more formidable challenge. These fake profiles will empower bad actors to infiltrate and manipulate online communities more easily; automation will reduce costs while increasing output quality. Fake accounts can be used to spread illicit or manipulative content, such as disinformation and hate speech, creating a skewed perception of popular opinions and norms[71]. We have already observed a large network of ChatGPT-driven fake profiles on Twitter/X[31]. These deceptive social bots engaged with each other through replies and retweets, created the false appearance of broad support for fake news sources and posted harmful comments while evading detection. Social media users are less likely to fact-check and more likely to share false claims when those claims appear to be liked or shared by many others[72]. This creates a significant vulnerability to manipulation by fake profiles. Potential harm could even extend to life-threatening matters. For instance, a strong correlation has been reported between the volume of online COVID-19 vaccine misinformation and a reduction in vaccination rates[73]. GenAI can be quickly weaponized to amplify such harm at scale.

3 Addressing the Threats

Addressing actuality-related challenges in LLMs requires a comprehensive strategy, as no single solution can fully mitigate the adverse consequences. Here, we outline various strategic dimensions that, when combined, may lead to more responsible and constructive technological utilization. Some solutions are technological and require building an entirely new LLM. Training a multi-billion-parameter LLM from scratch takes several months and hundreds of GPUs, which is beyond the reach of most academics. However, smaller models, such as those of LLaMA, Falcon, or Jais, are feasible in academia. For example, running a 7B model needs a single GPU, while a 13B model needs two GPUs.

Alignment and Safety: Safety and aligning LLMs with human values and intent have become a major concern for recent models like ChatGPT, LLaMA 2, and Jais. In fact, safety measures are increasingly being considered in all stages of chatbot development — data cleansing before training the base model, safety instruction-tuning[74], safety in the hidden prompt to the chatbot, and safety in the deployed chatbot via keywords and machine learning. The availability of open-source LLMs suggests that the effectiveness of alignment efforts, as well as other countermeasures, such as watermarking by large AI companies may be severely limited in mitigating the potential looming threats. Nevertheless, it remains crucial to make every conceivable effort that holds promise in restraining the counterproductive consequences of LLMs.

Modularized Knowledge Grounded Framework: One area where current LLMs significantly fall short is in producing timely, thorough, and well-organized presentations of factually dense information, such as situational and strategic reports. One way to ameliorate this shortcoming in the context of pre-trained LLMs involves a multi-step automated framework for gathering and organizing real-time event information. This modular design can create factually accurate content, which can be further refined using an LLM, as exemplified by SmartBook[75]. Initially developed for efficient ground-level reporting during the Russia-Ukraine conflict, SmartBook used LLMs to generate initial situational reports by streamlining event-based timelines, structured summaries, and consolidated references.

Retrieval-Augmented Generation: Retrieval-augmented generation (RAG)[76,77] incorporates contextual information from external sources into text generation. RAG mitigates the challenge of LLMs producing inaccurate content by enhancing their capabilities with external data. However, it requires efficient retrieval of grounded text at scale and robust evaluation.

Hallucination Control and Knowledge Editing: There are two types of LLM hallucinations[78]: (i) *faithfulness*, when the generated text is not faithful to the input context; and (ii) *factualness*, when the generated text is not factually correct with respect to world knowledge. Most recent attempts focus on solving the hallucination problem during the inference stage based on an LLM's self-consistency checking[79], cross-model verification[80,81], or checking against related knowledge[82]. The assumption is that an LLM has knowledge of a given concept, and sampled responses are likely to be similar and contain consistent facts. Another promising line of research focuses on opening up LLMs for knowledge editing. Factual errors can then be localized and fixed by injecting factual updates into the model[83–87]. Existing methods focus on factual updates for triples by precise editing. This can be extended to more complex knowledge representations, such as logical rules in the future. Another challenge is to evaluate the "ripple effects" of knowledge editing in language models. Current knowledge editing benchmarks check that a few paraphrases of the original fact are updated and some unrelated facts are untouched. More research must explore whether other facts logically derived from the edit are also changed accordingly.

Alleviating Exposure Bias: Exposure bias, i.e., favoring preexisting inductive biases over new ones, persists as a challenge in natural language generation, affecting the output quality of LLMs trained on fixed datasets[88]. Solutions such as selective upgrade, which dynamically derives relevant instruction-response pairs, aim to address this by improving the ability of LLMs to generalize effectively beyond their training data[89].

Better Evaluation: Existing evaluation methods such as BERTScore[90] and MoverScore[91] assume similar training and evaluation data distributions, which do not align with the evolving capabilities and requirements of LLMs. This discrepancy is particularly evident in scenarios involving zero-shot instructions and in-context learning, which prominently constitute disparate data distribution scenarios. Recently proposed evaluation measures such as GPTScore[64] and G-Eval[65] have shown reasonable correlations with human assessments in various tasks, including consistency, accuracy, and correctness. However, weak correlations (around 20–25%) remain in factuality assessments[64], suggesting that there is room for improvement. One potential direction is customizing factuality instructions for specific domains, such as medicine or law. Similar adjustments have demonstrated improved factuality assessment in the case of SelfCheckGPT[66], which is based on the idea that consistently replicable responses are rooted in factual accuracy, as opposed to those generated through stochastic sampling or hallucination, which tend to exhibit more variation.

Privacy and Data Protection: While organizations like OpenAI have implemented privacy controls and undergo periodic cybersecurity audits[92], user studies call for AI systems to adhere to additional data protection regulations, emphasizing steps like data anonymization, aggregation, and differential privacy[93]. While the European Union's AI Act[94] has transparency-related obligations based on identified risk levels, stronger regulation might facilitate access to third-party fact-verification APIs/plugins. Chatbot users could then opt-in to use such tools whenever verification is critical to the conversation.

Recognizing AI-Generated Content: LLM output is already nearly indistinguishable from human-written text[95]. For example, state-of-the-art tools to detect AI-generated content failed to distinguish between legitimate Twitter/X accounts and those managed through ChatGPT[31]. Furthermore, previous misinformation detectors[96] trained from automatically generated fake news training data still perform poorly on detecting human-generated or edited fake content[97]. Future models will likely narrow this gap further, making attempts to label AI-generated text unreliable. Watermarks will also be easy for malicious actors to bypass. Therefore, it is important to study multiple generators in multiple domains, for multiple languages, and using different detectors, and also to maintain constantly growing collections of up-to-date machine-generated content as in the M4 repository[95]. This issue parallels the rapid advancement of deepfake technology in various media formats, including fake videos, manipulated images, altered audio clips, and stylized text[98,99]. The emergence of sophisticated tactics, such as paraphrasing attacks aimed at evading detection[99] and adversarial perturbations resistant to image and video compression codecs[100], underscore the pressing need to address these concerns. Devising a robust and fool-proof detection technique, akin to a cat-and-mouse game, remains a formidable challenge.

Content Authenticity and Provenance: As AI-generated text becomes ubiquitous, fully human-written text may become more valuable. Technologies and standards for content authenticity and provenance already exist for video and image content[101]. These standards describe a way to cryptographically sign content so that the metadata around its creation can be proven not altered. We could use similar methods for text content to prove that they are not AI-generated. Since AI-generated content can cause harm when it spreads on social media, provenance proofs could be imposed to limit the spread of fake content before it has reached many people[28].

Regulation: Various regulatory efforts have emerged in light of the disruptive impact of emerging technologies. A new rule in China requires watermarks for AI-generated content, while ChatGPT was temporarily banned in Italy due to GDPR compliance concerns before being finally allowed after due compliance on transparency[102]. The European Union's AI Act is likely the first to regulate high-risk AI applications spanning multiple sectors[94]. In the USA, the Federal Trade Commission has issued warnings against creating misleading tools, emphasizing their prohibition under existing regulations and signaling the potential application of these rules to GenAI[103]. Along similar lines, the Canadian Directive on Automated Decision-Making has prescribed extensive guidelines that promote data-driven practices and federal compliance, transparency, and reduced negative algorithmic outcomes[104]. One of the challenges to the effectiveness of AI regulation stems from rapid technological advancements. On the one hand, controlling LLMs and their users can be as challenging as handling individuals engaged in phishing and misinformation. On the other hand, bad actors using open-source models will not be bound by regulation.

Public Education: Public awareness of deceptively "slick" content is essential, similar to our skepticism towards images doctored through Photoshop. This awareness extends to deepfake visual technology. A constructive way for experts to contribute to such education is through resources such as tutorial videos and code, explaining the basic technology behind

ChatGPT. However, raising awareness about LLM-based chatbots is challenging, as the risks they pose demand immediate attention, unlike the gradual understanding of deepfakes for visual content that has developed over the last few years. Another caveat is that skeptical citizens may lose trust in credible, authoritative sources of information, and thus become more vulnerable to conspiracy theories.

4 Fact-Checking Opportunities

Leveraging LLMs to foster factuality is as important as addressing the challenges of LLMs. In fact, one could argue that verifying the truthfulness of a claim may be more important than detecting whether it is AI-generated. LLMs present several promising opportunities for fact-checkers and journalists, some of which are highlighted below. It is important, however, to remain vigilant about the challenges and the ethical considerations associated with such applications of LLMs, including bias, transparency, and the potential for misuse. Balancing the advantages of LLMs with responsible and ethical practices is essential to harness their full potential in domains such as fact-checking.

Fact-Checking Support: Although GenAI models lack a concept of truth, they can be used to assist fact-checkers in verifying claims. LLMs can transcribe speeches, debates, interviews, online videos, and news broadcasts, summarize extensive documents, and help create concise lists of crucial claims[105–107]. These claims can be further organized into fine-grained claim frames, and the claim-claim relations can be utilized for cross-media cross-lingual fact checking[108] and factual error correction[109]. This capability is particularly valuable for processing large volumes of news content, enabling fact-checkers to monitor a vast amount of potentially misleading information. After a fact-checked article has been published, the same claims often reappear. LLMs can help identify sections of documents that repeat a previously fact-checked claim or that make a claim semantically equivalent to a previously verified one. This task does not involve determining the factuality or the falsehood of the claim, as this was already established by fact-checkers or journalists[43]. Such promising opportunities for LLMs to support fact-checking come with some caveats. First, there are risks of errors when summarizing long documents, such as omitting important context or inserting plausible text that is outside the source material. Second, we need to understand how humans will interact with AI-aided fact-checking. A recent study found that fact-checks generated by ChatGPT, even when accurate, did not significantly affect participants' ability to discern headline accuracy or to share accurate news[33]. Worse, they were harmful in specific cases, such as increasing beliefs in certain false claims. These findings highlight the importance of evaluating unintended consequences of GenAI applications.

Stance Detection: There are indications that ChatGPT and other LLMs can help with downstream fact-checking tasks, such as stance detection[110]. On the other hand, whether the LLM has seen the test data, as it was trained on many publicly available datasets remains as a partinent question. Recent studies suggest that ChatGPT is more inclined to excel in stance detection tasks compared to tasks involving emotions or pragmatic analyses, even though it might not have undergone pre-training on evaluation sets explicitly tailored for stance detection tasks[111].

Domain-Specific Chat Support: There have been attempts to use LLMs to provide a chatbot interface to a domain-specific controlled corpus[112]. Organizations could provide such a service to allow users to search for information on a collection of factually verified articles. Some caution is needed, as LLMs can insert plausible but incorrect information even with a controlled corpus.

5 Conclusion

Any tool capable of generating novel content also has the potential to produce misleading content[113]. Therefore, anyone using LLMs to compose news articles, academic reports, emails, or any text must verify even the most basic facts, regardless of how fluent the text appears. Given the rapid and widespread growth in the use of LLMs, society must act quickly with appropriate regulation, education, and collaboration. Below, we propose an urgent agenda for individuals, governments, and democratic societies.

Coordination and Collaboration: Towards ensuring the responsible development and deployment of GenAI technologies, nations must make coordinated research investments and establish infrastructure capable of dynamically adapting legislative safeguards, much like the measures in place for other groundbreaking technologies such as human germline editing[114]. Moreover, fostering collaborations between political actors and industry leaders is essential to prevent an arms race between AI and AI-detection tools. By coordinating these efforts globally, we can promote transparent, constructive, and responsible technological development, ensuring that GenAI benefits humanity while minimizing potential risks and abuses.

Regulations: Key components of comprehensive regulations in the realm of GenAI include stringent law enforcement to mitigate intentional or inadvertent harm from technology use[115], regulatory frameworks for high-risk tech production and sales[103], and industry-wide standards for ethical GenAI usage. Special attention should be paid to creating guidelines for journalists who use GenAI and for labeling public ads generated by AI. Adopting standardized evaluation measures tailored to critical sectors will also be vital to ensuring that technology benefits society while minimizing the risks[48,64].

Promoting AI Literacy: To promote AI literacy globally, we recommend three key actions: (i) conducting AI literacy programs for people of all ages[100], (ii) incorporating AI education with an emphasis on ethics into the graduate-level curricula[116], and (iii) sensitizing digital consumers to the potential harms and root causes of GenAI[117]. These measures will empower individuals to engage with AI responsibly and safely, fostering a more knowledgeable and aware global community.

Technological Development: As common R&D best practices, we recommend integrating clear and accessible informative material for users upfront to clarify the limits (e.g., in terms of factuality) and the risks (e.g., because of the "authoritative" tone) of LLMs. To enhance safety, robust guardrails should be implemented in LLM-based conversations[118]. Chatbots should be equipped with evidence-supporting capabilities[119] to bolster credibility, while external knowledge sources like retrieval-augmented LLMs and knowledge graphs can ensure factual consistency[120]. Lastly, the development of coordination detection algorithms is essential to identify suspiciously similar and potentially harmful content[121] generated by malicious actors using LLMs[28].

Author Contributions

All the authors contributed to conceptualizing, preparing, and finalizing the manuscript. The author names are arranged alphabetically by the last names.

Funding Information

M.C. is supported by the Institute for Basic Science (grant IBS-R029-C2) and the National Research Foundation of Korea (grant RS-2022-00165347). T.C. acknowledges the financial support of Wipro AI. I.A. is supported in part by the European Union (ERC, ExplainYourself, 101077481). G.L.C. is supported by the National Science Foundation (grant 2239194). E.F. and F.M. are partly supported by DARPA (grant HR001121C0169). F.M. is also partly supported by the Knight Foundation and Craig Newmark Philanthropies. G.Z.'s fact-checking project receives funding from the European Union through multiple grants and is part of Meta's 3PFC Program. H.J.'s research is partially supported by U.S. DARPA SemaFor Program No. HR001120C0123. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

Competing Interests

The authors declare no competing interests.

Additional Information

Materials & Correspondence should be emailed to Tanmoy Chakraborty (tanchak@iitd.ac.in).

References

1. Shannon, C. E. A mathematical theory of communication. *The Bell Syst. Tech. J.* **27**, 379–423, DOI: 10.1002/j.1538-7305.1948.tb01338.x (1948).
2. Huang, J. & Chang, K. C.-C. Towards Reasoning in Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, 1049–1065, DOI: 10.18653/v1/2023.findings-acl.67 (Association for Computational Linguistics, Toronto, Canada, 2023).
3. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. *et al.* Improving language understanding by generative pre-training (2018).
4. OpenAI. GPT-4 technical report. *arXiv:2303.08774* (2023).
5. Touvron, H. *et al.* Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
6. Introducing ChatGPT — openai.com. https://openai.com/blog/chatgpt. [Accessed 27-09-2023].
7. Hu, K. ChatGPT sets record for fastest-growing user base - analyst note. *Reuters* (2023).
8. Thoppilan, R. *et al.* Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).

9. Taori, R. *et al.* Stanford Alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca (2023).

10. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. https://lmsys.org/blog/2023-03-30-vicuna/. [Accessed 15-09-2023].

11. Introducing Claude. https://www.anthropic.com/index/introducing-claude. [Accessed 15-09-2023].

12. Falcon. https://falconllm.tii.ae/falcon.html. [Accessed 15-09-2023].

13. Jurassic-2 models. https://docs.ai21.com/docs/jurassic-2-models. [Accessed 20-09-2023].

14. Sengupta, N. *et al.* Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint 2308.16149* (2023). 2308.16149.

15. Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).

16. Brown, T. *et al.* Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, vol. 33, 1877–1901 (Curran Associates, Inc., 2020).

17. Zhao, W. X. *et al.* A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

18. Li, S. *et al.* Open-domain hierarchical event schema induction by incremental prompting and verification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5677–5697, DOI: 10.18653/v1/2023.acl-long.312 (Association for Computational Linguistics, Toronto, Canada, 2023).

19. Wei, J. *et al.* Emergent abilities of large language models. *Trans. Mach. Learn. Res.* **2022** (2022).

20. Bang, Y. *et al.* A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. Preprint 2302.04023, arXiv (2023). DOI: 10.48550/arXiv.2302.04023.

21. Bergstrom, C. T. & Ogbunu, C. B. ChatGPT isn't 'hallucinating.' It's Bullshitting. https://undark.org/2023/04/06/chatgpt-isnt-hallucinating-its-bullshitting/ (2023). Accessed 2 October 2023.

22. Sison, A. J. G., Daza, M. T., Gozalo-Brizuela, R. & Garrido-Merchán, E. C. ChatGPT: More Than a "Weapon of Mass Deception" – Ethical Challenges and Responses from the Human-Centered Artificial Intelligence (HCAI) Perspective. *Int. J. Human-Computer Interact.* **0**, 1–20, DOI: 10.1080/10447318.2023.2225931 (2023). https://doi.org/10.1080/10447318.2023.2225931.

23. Iftikhar, L. *et al.* Docgpt: Impact of chatgpt-3 on health services as a virtual doctor. *EC Paediatr.* **12**, 45–55 (2023).

24. Chin, H. *et al.* User-chatbot conversations during the COVID-19 pandemic: Study based on topic modeling and sentiment analysis. *J. Med. Internet Res.* **25**, e40922, DOI: 10.2196/40922 (2023).

25. Palanica, A., Flaschner, P., Thommandram, A., Li, M. & Fossat, Y. Physicians' perceptions of chatbots in health care: Cross-sectional web-based survey. *J Med Internet Res* **21**, e12887, DOI: 10.2196/12887 (2019).

26. Peskoff, D. & Stewart, B. Credible without credit: Domain experts assess generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 427–438, DOI: 10.18653/v1/2023.acl-short.37 (Association for Computational Linguistics, Toronto, Canada, 2023).

27. Srivastava, B. Did chatbots miss their "Apollo moment"? Potential, gaps, and lessons from using collaboration assistants during COVID-19. *Patterns* **2**, 100308, DOI: https://doi.org/10.1016/j.patter.2021.100308 (2021).

28. Menczer, F., Crandall, D., Ahn, Y.-Y. & Kapadia, A. Addressing the harms of AI-generated inauthentic content. *Nat. Mach. Intell.* **5**, 678–680, DOI: 10.1038/s42256-023-00690-w (2023).

29. Patel, A. & Sattler, J. Creatively malicious prompt engineering. Tech. Rep., WithSecure Labs (2023).

30. Brewster, J., Wang, M. & Palmer, C. Plagiarism-Bot? How Low-Quality Websites Are Using AI to Deceptively Rewrite Content from Mainstream News Outlets. https://www.newsguardtech.com/misinformation-monitor/august-2023/ (2023). [Accessed 23 Sep 2023].

31. Yang, K.-C. & Menczer, F. Anatomy of an AI-powered malicious social botnet. Preprint 2307.16336, arXiv (2023). DOI: 10.48550/arXiv.2307.16336.

32. Verma, P. & Oremus, W. ChatGPT invented a sexual harassment scandal and named a real law prof as the accused. *Wash. Post* (2023).

33. DeVerna, M. R., Yan, H. Y., Yang, K.-C. & Menczer, F. Artificial intelligence is ineffective and potentially harmful for fact checking. Preprint 2308.10800, arXiv (2023). DOI: 10.48550/arXiv.2308.10800.

34. Ferrara, E. The history of digital spam. *Commun. ACM* **62**, 82–91, DOI: 10.1145/3299768 (2019).

35. Vincent, J. Google's AI chatbot Bard makes factual error in first demo (2023).

36. Marcus, G. Deep Learning Is Hitting a Wall (2022).

37. Ferrara, E. Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. *Should ChatGPT be Biased? Challenges Risks Bias Large Lang. Model.* (2023).

38. Pan, Y. *et al.* On the risk of misinformation pollution with large language models. *arXiv preprint 2305.13661* (2023). 2305.13661.

39. Dall.e 2. https://openai.com/dall-e-2. [Accessed 15-09-2023].

40. Midjourney. https://www.midjourney.com/. [Accessed 15-09-2023].

41. Stable diffusion. https://stablediffusionweb.com/. [Accessed 15-09-2023].

42. Mirsky, Y. & Lee, W. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* **54**, DOI: 10.1145/3425780 (2021).

43. CLEF2022-CheckThat! — sites.google.com. https://sites.google.com/view/clef2022-checkthat.

44. Liu, N. F., Zhang, T. & Liang, P. Evaluating verifiability in generative search engines (2023). 2304.09848.

45. Ji, Z. *et al.* Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, DOI: 10.1145/3571730 (2023).

46. Vincent, J. AI-generated answers temporarily banned on coding Q&A site Stack Overflow — theverge.com. https://www.theverge.com/2022/12/5/23493932/chatgpt-ai-generated-answers-temporarily-banned-stack-overflow-llms-dangers. [Accessed 08-09-2023].

47. Kusunose, K., Kashima, S. & Sata, M. Evaluation of the accuracy of ChatGPT in answering clinical questions on the Japanese Society of Hypertension Guidelines. *Circ. J.* **87**, 1030–1033, DOI: 10.1253/circj.CJ-23-0308 (2023).

48. Antaki, F., Touma, S., Milad, D., El-Khoury, J. & Duval, R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. *Ophthalmol. Sci.* **3**, 100324, DOI: https://doi.org/10.1016/j.xops.2023.100324 (2023).

49. Abels, G. Can ChatGPT fact-check? We tested. - Poynter — poynter.org. https://www.poynter.org/fact-checking/2023/chatgpt-ai-replace-fact-checking/. [Accessed 27-09-2023].

50. Bai, H., Voelkel, J. G., Eichstaedt, j. C. & Willer, R. Artificial intelligence can persuade humans on political issues, DOI: 10.31219/osf.io/stakv (2023).

51. Brashier, N. M. & Marsh, E. J. Judging truth. *Annu. review psychology* **71**, 499–515 (2020).

52. IFCN fact-checking organizations on WhatsApp. https://faq.whatsapp.com/5059120540855664 (2023). Accessed 2 October 2023.

53. Running LLaMA 7B and 13B on a 64GB M2 MacBook Pro with llama.cpp — til.simonwillison.net. https://til.simonwillison.net/llms/llama-7b-m2. [Accessed 08-09-2023].

54. Falcon LLM — falconllm.tii.ae. https://falconllm.tii.ae/falcon-180b.html. [Accessed 08-09-2023].

55. Volunteering, C. Halo effect — wikipedia.org. https://en.wikipedia.org/wiki/Halo_effect. [Accessed 27-09-2023].

56. OpenAI. Lawyer used ChatGPT in court—and cited fake cases. A judge is considering sanctions. https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering [Accessed 15-09-2023].

57. Chakraborty, T. & Masud, S. Judging the creative prowess of AI. *Nat. Mach. Intell.* 1–1 (2023).

58. Srivastava, A. *et al.* Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Mach. Learn. Res.* (2023).

59. Wang, A. *et al.* GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355, DOI: 10.18653/v1/W18-5446 (Association for Computational Linguistics, Brussels, Belgium, 2018).

60. Wang, A. *et al.* SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 294 (Curran Associates Inc., Red Hook, NY, USA, 2019).

61. Lin, S., Hilton, J. & Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252, DOI: 10.18653/v1/2022.acl-long.229 (Association for Computational Linguistics, Dublin, Ireland, 2022).

62. Min, S. *et al.* FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *arXiv preprint 2305.14251* (2023). 2305.14251.

63. Golchin, S. & Surdeanu, M. Time Travel in LLMs: Tracing Data Contamination in Large Language Models (2023). 2308.08493.

64. Fu, J., Ng, S.-K., Jiang, Z. & Liu, P. GPTScore: Evaluate as you desire (2023). 2302.04166.

65. Liu, Y. *et al.* G-Eval: NLG evaluation using GPT-4 with better human alignment (2023). 2303.16634.

66. Manakul, P., Liusie, A. & Gales, M. J. F. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models (2023). 2303.08896.

67. Wang, P. *et al.* Large language models are not fair evaluators (2023). 2305.17926.

68. Lee, N., Bang, Y., Madotto, A. & Fung, P. Towards few-shot fact-checking via perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1971–1981, DOI: 10.18653/v1/2021.naacl-main.158 (Association for Computational Linguistics, Online, 2021).

69. Coles, C. 11% of data employees paste into ChatGPT is confidential - Cyberhaven — cyberhaven.com. https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt/. [Accessed 08-09-2023].

70. Meta's Third-Party Fact-Checking Program | Meta Journalism Project — facebook.com. https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking.

71. Truong, B. T., Lou, X., Flammini, A. & Menczer, F. Vulnerabilities of the online public square to manipulation. Preprint 1907.06130, arXiv (2023). DOI: 10.48550/arXiv.1907.06130.

72. Avram, M., Micallef, N., Patil, S. & Menczer, F. Exposure to social engagement metrics increases vulnerability to misinformation. *HKS Misinformation Rev.* **1**, DOI: 10.37016/mr-2020-033 (2020).

73. Pierri, F. *et al.* Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Sci. Reports* **12**, 5966, DOI: 10.1038/s41598-022-10070-w (2022).

74. Wang, Y., Li, H., Han, X., Nakov, P. & Baldwin, T. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint 2308.13387* (2023). 2308.13387.

75. Reddy, R. G. *et al.* Smartbook: AI-assisted situation report generation (2023). 2303.14337.

76. Yu, W. *et al.* A survey of knowledge-enhanced text generation. *ACM Comput. Surv.* **54**, DOI: 10.1145/3512467 (2022).

77. Guu, K., Lee, K., Tung, Z., Pasupat, P. & Chang, M.-W. Realm: retrieval-augmented language model pre-training. *ArXiv* (2020).

78. Filippova, K. Controlled hallucinations: learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 864–870, DOI: 10.18653/v1/2020.findings-emnlp.76 (Association for Computational Linguistics, Online, 2020).

79. Gou, Z. *et al.* Critic: large language models can self-correct with tool-interactive critiquing. In *arxiv* (2023).

80. Cohen, R., Hamri, M., Geva, M. & Globerson, A. Lm vs lm: detecting factual errors via cross examination. In *arxiv* (2023).

81. Du, Y., Li, S., Torralba, A., Tenenbaum, J. B. & Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *arxiv* (2023).

82. Dziri, N., Madotto, A., Zaïane, O. & Bose, A. J. Neural path hunter: reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2197–2214, DOI: 10.18653/v1/2021.emnlp-main.168 (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021).

83. De Cao, N., Aziz, W. & Titov, I. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6491–6506, DOI: 10.18653/v1/2021.emnlp-main.522 (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021).

84. Mitchell, E., Lin, C., Bosselut, A., Finn, C. & Manning, C. D. Fast model editing at scale. *arXiv preprint arXiv:2110.11309* (2021).

85. Mitchell, E., Lin, C., Bosselut, A., Manning, C. D. & Finn, C. Memory-based model editing at scale. In *International Conference on Machine Learning*, 15817–15831 (PMLR, 2022).

86. Meng, K., Bau, D., Andonian, A. & Belinkov, Y. Locating and editing factual associations in gpt. *Adv. Neural Inf. Process. Syst.* **35**, 17359–17372 (2022).

87. Meng, K., Sharma, A. S., Andonian, A., Belinkov, Y. & Bau, D. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229* (2022).

88. Schmidt, F. Generalization in Generation: A closer look at Exposure Bias. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, 157–167, DOI: 10.18653/v1/D19-5616 (Association for Computational Linguistics, Hong Kong, 2019).

89. Yu, P. & Ji, H. Self information update for large language models through mitigating exposure bias. *arXiv preprint arXiv:2305.18582* (2023).

90. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020* (OpenReview.net, 2020).

91. Zhao, W. *et al.* MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 563–578, DOI: 10.18653/v1/D19-1053 (Association for Computational Linguistics, Hong Kong, China, 2019).

92. OpenAI. Security & privacy. https://openai.com/security. [Accessed 27-09-2023].

93. Sebastian, G. Privacy and data protection in chatgpt and other ai chatbots: Strategies for securing user information. *Int. J. Secur. Priv. Pervasive Comput.* **15**, 1–14, DOI: 10.4018/IJSPPC.325475 (2023).

94. EUR-Lex - 52021PC0206 - EN - EUR-Lex — eur-lex.europa.eu. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021

95. Wang, Y. *et al.* M4: Multi-generator, Multi-domain, and Multi-lingual Black-Box Machine-Generated Text Detection. *arXiv:2305.14902* (2023).

96. Fung, Y. *et al.* Infosurgeon: cross-media fine-grained information consistency checking for fake news detection. In *Proc. The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)* (2021).

97. Huang, K.-H., McKeown, K., Nakov, P., Choi, Y. & Ji, H. Faking fake news for real fake news detection: propaganda-loaded training data generation. In *Proc. The 61st Annual Meeting of the Association for Computational Linguistics (ACL2023) Findings* (2023).

98. Groh, M. *et al.* Human detection of political speech deepfakes across transcripts, audio, and video (2023). 2202.12883.

99. Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W. & Feizi, S. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156* (2023).

100. Hussain, S., Neekhara, P., Jere, M., Koushanfar, F. & McAuley, J. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 3348–3357 (2021).

101. Content Authenticity Initiative — contentauthenticity.org. https://contentauthenticity.org/. [Accessed 08-09-2023].

102. ChatGPT: OpenAI reopens the platform in Italy guaranteeing more transparency and more rights to European users and non-users — tbs-sct.canada.ca. https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9881490. [Accessed 27-09-2023].

103. Chatbots, deepfakes, and voice clones: AI deception for sale — ftc.gov. https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale. [Accessed 08-09-2023].

104. Directive on Automated Decision-Making — tbs-sct.canada.ca. https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592. [Accessed 27-09-2023].

105. Gangi Reddy, R. *et al.* NewsClaims: A new benchmark for claim detection from news with attribute knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 6002–6018, DOI: 10.18653/v1/2022.emnlp-main.403 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022).

106. Sundriyal, M., Kulkarni, A., Pulastya, V., Akhtar, M. S. & Chakraborty, T. Empowering the fact-checkers! automatic identification of claim spans on Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7701–7715, DOI: 10.18653/v1/2022.emnlp-main.525 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022).

107. Sundriyal, M., Singh, P., Akhtar, M. S., Sengupta, S. & Chakraborty, T. Desyr: definition and syntactic representation based claim detection on the web. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 1764–1773 (2021).

108. Huang, K.-H., Zhai, C. & Ji, H. CONCRETE: Improving cross-lingual fact-checking with cross-lingual retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*, 1024–1035 (International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022).

109. Huang, K.-H., Chan, H. P. & Ji, H. Zero-shot faithful factual error correction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5660–5676, DOI: 10.18653/v1/2023.acl-long.311 (Association for Computational Linguistics, Toronto, Canada, 2023).

110. Zhang, B., Ding, D. & Jing, L. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548* (2022).

111. Kocoń, J. *et al.* ChatGPT: Jack of all trades, master of none. *Inf. Fusion* **99**, 101861, DOI: https://doi.org/10.1016/j.inffus.2023.101861 (2023).

112. Shankar, A. Remembering Conversations: Building Chatbots with Short and Long-Term Memory on AWS — itnext.io. https://itnext.io/remembering-conversations-building-chatbots-with-short-and-long-term-memory-on-aws-c1361c130046. [Accessed 28-09-2023].

113. Ferrara, E. GenAI Against Humanity: Nefarious Applications of Generative Artificial Intelligence and Large Language Models. *arXiv preprint arXiv:2310.00737* (2023).

114. Login | The National Academies Press — nap.nationalacademies.org. https://nap.nationalacademies.org/download/24623. [Accessed 08-09-2023].

115. McCallum, S. ChatGPT banned in Italy over privacy concerns. https://www.bbc.com/news/technology-65139406. [Accessed 08-09-2023].

116. New UK university principles promote AI literacy and integrity — universityworldnews.com. https://www.universityworldnews.com/post.php?story=20230704155107330. [Accessed 08-09-2023].

117. Department for Education. Generative artificial intelligence in education departmental statement. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1146540/Generative_artificial_intel [Accessed 08-09-2023].

118. Cohen, J. Right on Track: NVIDIA Open-Source Software Helps Developers Add Guardrails to AI Chatbots — blogs.nvidia.com. https://blogs.nvidia.com/blog/2023/04/25/ai-chatbot-guardrails-nemo/. [Accessed 08-09-2023].

119. Chen, A. & Chen, D. O. Accuracy of chatbots in citing journal articles. *JAMA Netw Open* **6**, e2327647 (2023).

120. Spataro, J. Introducing Microsoft 365 Copilot – your copilot for work - The Official Microsoft Blog — blogs.microsoft.com. https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/. [Accessed 08-09-2023].

121. Pacheco, D. *et al.* Uncovering coordinated networks on social media: Methods and case studies. In *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, vol. 15, 455–466, DOI: 10.1609/icwsm.v15i1.18075 (2021).