Misinformation and Disinformation in Generative AI—A Survey

Ramya Srinivasan(✉)

Fujitsu Research of America, Santa Clara, USA
ramya@fujitsu.com

Abstract. Recent years have witnessed the rapid growth and adoption of generative AI in diverse industries such as healthcare, retail, financial services, and more. In the midst of these advancements, there are also rising concerns about misinformation, disinformation, and other ethical issues surrounding generative AI. Towards addressing these concerns, a variety of countermeasures are being developed. In this survey paper, we provide a comprehensive overview of various research methods and best practices to detect and mitigate misinformation and disinformation associated with generative AI, in particular large language models (LLMs). Specifically, we provide a holistic summary of related datasets and benchmarks, detection and mitigation methods, review human-centric best practices and guidelines suggested towards addressing these challenges, and outline open problems and future directions. We hope that the survey will help researchers and practitioners understand the latest advances at the intersection of generative AI and misinformation/disinformation.

Keywords: misinformation · disinformation · LLMs · generative AI

1 Introduction

Large Language Models (LLMs), and more broadly generative AI technologies are becoming ubiquitous—from entertainment and education to marketing and medicine, LLMs such as GPT, LLaMa, PaLM, Claude, etc. are being deployed across a variety of applications [123]. Generative AI technologies, have in many ways, ushered in a new era of AI advancements, transforming the way people can search and interact with technology [4,52,71].

As with any technology, there can also be downsides with generative AI technologies. Researchers, regulators, and other stakeholders have highlighted some of the adverse impacts of generative AI on people and society. For example, in a recent work [39], the authors delve into the issue concerning proliferation of dubious information stemming from the use of LLMs. The authors note that misinformation (i.e., unintended sharing of false data) and disinformation (i.e., deliberate spread of misleading content to further specific agendas) pose a major challenge in information creation, dissemination, and consumption [39]. Studies

X. Wu et al. (Eds.): PAKDD 2025, LNAI 15873, pp. 290–307, 2025.
https://doi.org/10.1007/978-981-96-8183-9_23

have also shown that LLMs are vulnerable to user attacks owing to their human-like reasoning capabilities [110]. Additionally, researchers argue that the use of synthetic data generated from LLMs pose major risks in terms of circumventing user consent and could also compromise on dataset diversity and representation [98]. Several works have also discussed issues related to copyright compliance [67], and the potential marginalization of human creativity due to the overuse of LLMs [109].

Researchers argue that current methods for evaluating LLM safety are inadequate, as they are largely model-centric and do not take into account the broader contexts in which these tools are used [77]. Somewhat corroborating this issue, the authors in [48] present results from semi-structured interviews with government, legal, and civil society stakeholders who inform policy and advocacy around responsible AI efforts, to highlight the challenges faced by these stakeholders. Furthermore, recent research has also shown that existing AI red-teaming practices diverge along several axes, such as vagueness in the purpose of the activity, in the artifact under evaluation, in the setting in which the activity is conducted (e.g., actors, resources, and methods) and the resulting decisions it informs (e.g., reporting, disclosure, and mitigation) [25]. In a similar vein, the authors in [99] argue that significant and understudied harms originate from differing practices of transparency and accountability in the open source LLM community [99].

Despite concerns such as those listed above, researchers posit that the benefits of LLM outweigh the risks and that through appropriate countermeasures, some of these risks can be alleviated or minimized [16,23,47]. Unlike previous related surveys that have focused on LLM architectures [123], LLM applications in specific domains [55,105,110], LLM safety evaluation [25], biases in LLMs [26], and misinformation in online social media [9], the objective of this survey paper is to provide a comprehensive overview of recent work related to misinformation and disinformation detection and mitigation in the context of generative AI- in particular LLMs. Unlike previous machine learning approaches designed to detect and mitigate misinformation [9], the use of Generative AI poses new challenges and opportunities both in terms of fueling the problem and in terms of offering new solutions [63]. This calls for a review of recent progress in LLM based misinformation and disinformation detection and mitigation methods. We review related datasets and benchmarks, summarize best practices and guidelines suggested to address some of the existing concerns, and outline open challenges and future directions.

2 Survey Methodology

For this survey, we reviewed papers from conferences spanning a range of academic communities such as computational linguistics (ACL, NAACL, EMNLP, COLM), data mining (ACM KDD, PAKDD), machine learning (ICML, ICLR, NeurIPS, IJCAI), computer vision (CVPR), knowledge management (CIKM), and AI ethics (ACM FAccT, AAAI AIES). We focused on papers published in

these conferences over the last three years since there has been a surge in generative AI based applications during this time frame. We also included relevant articles from journals. The survey results are organized along different themes –datasets and benchmarks, misinformation and disinformation detection and mitigation methods, and best practices and guidelines.

As noted in Sect. 1, spread of dubious information from LLMs could be either intentional disinformation or unintentional misinformation. We provide a fine-grained analysis of different sub-types of misinformation and disinformation as outlined in Table 1. Disinformation could manifest in various forms such as fake news, fake accounts, fake videos etc., thus spanning multiple data modalities. Misinformation primarily occurs as hallucinations and incorrect information.

Table 1. Datasets and benchmarks related to LLM based misinformation and disinformation detection and mitigation. Note: As the survey focuses on the specific context of LLMs, this list therefore does not include non-generative AI related works

Type	Issue	Datasets/Benchmarks
Misinformation	Hallucinations	[41,53,58,70,74,75,87] [21,27,115]
Misinformation	Ambiguous statements and factual errors	[12,36,66,104,106,115,120,128] [19,35,90,93]
Disinformation	Fake News, Rumors	[8,15,37,38,54,85]
Disinformation	Deepfakes, Forgery	[40,50,68,76,96,108,122], [20,45,92,107,117]

3 Misinformation: Detection and Mitigation Methods

In this section, different types of misinformation are outlined along with corresponding detection and mitigation methods, related datasets, and benchmarks.

3.1 Hallucinations

Hallucinations are a prominent type of misinformation in LLMs. Hallucinations can be thought of as responses that are incongruent to the prompts, often characterized by non-sensical, imaginary, and counterfactual statements. Research has shown that hallucinations occur due to a variety of reasons such as the output of LLMs being not constrained to be synonymous with claims for which they have evidence [10], due to the use of specific pre-training and inference methods [53], due to the presence of spurious data [33], etc. Despite some arguments suggesting that hallucinations could be associated with coherent narrative generation and fact verification [31,81,84,88], by and large, it is argued that hallucinations not only compromise on the utility of LLMs, but also their safety [10,119]. As a result, there is a growing body of work analyzing the causes of hallucinations, and ways to detect and mitigate them [42].

Detection and Mitigation Methods. In [111], the authors propose an approach to detect hallucinations by prompting the model multiple times to reconstruct the input query using the generated answer and subsequently, by quantifying the inconsistency level between the original query and the reconstructed queries. In [53], the authors propose to detect hallucinations based on extracting factual statements from responses and judging the trustfulness of the generations. The authors in [82] demonstrate that tokens preceding a hallucination can already predict the subsequent hallucination even before it occurs. In [114], the authors propose two methods to mitigate multimodal hallucinations–one a training objective that enables the model to reduce hallucinations by learning from regular instruction data, and two, a data filtering strategy to prevent harmful training data from exacerbating model hallucinations. In [125], a training free method called Residual Visual Decoding is proposed to counteract the issue of multimodal snowballing. In [124], a comprehensive analysis of hallucination in simultaneous machine translation is conducted. In [2], the authors comprehensively review knowledge-graph-based augmentation techniques in LLMs, focusing on their efficacy in mitigating hallucinations.

The authors in [103] propose HAllucination Diversity-Aware Sampling (referred to as HADAS) to select diverse hallucinations for annotations in active learning for LLM finetuning. In [72], a new metric called ALOHa, which leverages large language models (LLMs) to measure object hallucinations is proposed. In [80], the authors present context-aware decoding (CAD), which follows a contrastive output distribution that amplifies the difference between the output probabilities when a model is used with and without context to mitigate hallucinations in LLMs. In [95], the authors demonstrate the feasibility of mitigating hallucinations by verifying and minimizing the inconsistency between external knowledge present in the alignment data and the intrinsic knowledge embedded within foundation LLMs. In [113], the authors propose a novel and effective method to mitigate false premise hallucinations by constraining the false premise attention heads during the model inference process.

In [3], the authors introduce HalluMeasure, a new LLM-based hallucination detection mechanism that decomposes an LLM response into atomic claims, and evaluates each atomic claim against the provided reference context. In [18], the authors propose an autonomous LLM-based agent framework, called HaluAgent, which enables relatively smaller LLMs (e.g. Baichuan2-Chat 7B) to actively select suitable tools for detecting hallucination across text, code, and mathematical expression. In [60], the authors present an in-depth investigation into the object hallucination problem specifically within the CLIP model.

Datasets and Benchmarks. A novel token-level, reference-free hallucination detection task and an associated annotated dataset named HaDeS (HAllucination DEtection dataSet) is provided in [58], to aid in detecting hallucinations in scenarios where ground truth data may not be available. In [74], the authors develop a novel metric, which they call mFACT, to evaluate the faithfulness of LLM generated non-English summaries. In [41], the authors present ANAH, a

bilingual dataset that offers ANalytical Annotation of Hallucinations in LLMs within Generative Question Answering. In [70], RAGTruth, a corpus tailored for analyzing word-level hallucinations in various domains and tasks within the standard RAG frameworks for LLM applications is presented.

The authors in [53] propose a benchmark called HaluEval 2.0 that contains 8,770 questions from five domains including biomedicine, finance, science, education, and open domain. In [13], a novel visual dialogue hallucination evaluation benchmark is proposed consisting of samples with five-turn questions about an edited image and its original version. In [75], a dataset to study hallucinations in LLM based summarizations is provided. In another related work [87], the authors propose a new evaluation benchmark on topic-focused dialogue summarization, generated by LLMs of varying sizes. In [27], the authors conduct the first systematic analysis of the effect of fine-grained object grounding on large vision language model (LVLM) hallucination under an evaluation protocol that more realistically captures LVLM hallucination. Taking a step further, in [21], the authors introduce SoraDetector, a novel unified framework designed to detect hallucinations across diverse large text to video models and introduce an associated meta-evaluation benchmark called T2VHaluBench.

3.2 Factual Errors and Ambiguous Information

Ambiguous and incorrect information are a type of seemingly valid arguments made by LLMs supporting certain viewpoints. Such arguments could be deliberate (e.g., fake news, cherry picking, red herring, etc.) in which case they constitute disinformation (detailed in the subsequent section), but some such arguments might be unintentional arising due to the limited reasoning capabilities or lack of knowledge of LLMs. For example, circular reasoning wherein the LLMs uses the premise of the statement itself to reason could be one type of unintentional misinformation. The statement 'She is the best because she is better than anyone else' is an illustration to the point [5]. Factual errors are another type of unintentional false information [115]. Below, we summarizes the methods and benchmarks associated with such misinformation types.

Detection and Mitigation Methods. In [65], a human-in-the-loop evaluation framework for fact-checking novel misinformation claims is proposed for the use-case of COVID-19 treatments by identifying social media messages that support them. In [28], the authors propose a method for generating synthetic data by annotating diverse model-generated summaries using a LLM for verifying factual consistencies. In [102], the authors highlight the crucial role of qualitative causal structure in characterizing and verifying scientific claims based on evidence. Investigating the cause for falsity in LLMs, the authors in [32] study the topic of harmful imitation through the lens of a model's internal representations, and identify two related phenomena: overthinking and false induction heads as potential causes of harmful imitation.

Noting that temporal misalignment could be a source of factual inconsistencies, in [118], the authors propose a method to predict how long a given fact

will remain true, which in turn can help in identifying which facts are prone to rapid change and can help models avoid reciting outdated information. In related works such as [24,30], the authors study the sources of data contamination in big datasets uncovering issues such as data redundancy and low quality among several others as potential causes of misinformation and disinformation. In [22], the authors demonstrate that factuality and reasoning in LLMs can be enhanced through multi-agent debate.

In [97], the authors propose a novel metric called —MOdel kNowledge relIabiliTy scORe (MONITOR)— to directly measure LLMs' factual reliability. In [43], the authors investigate the phenomenon of LLMs possessing correct answer knowledge yet provide incorrect response from the perspective of inference dynamics, by identifying the factual questions that query the same triplet knowledge but result in different answers. Investigating what influences users' re-posting behavior of misinformation on social media, in [59], the authors propose a computational approach to model users' latent susceptibility levels based on various factors such as their demographic background and political ideology.

Datasets and Benchmarks. In [115], the authors propose AlignScore, a new holistic metric that applies to a variety of factual inconsistency scenarios and further develop a unified training framework of the alignment function by integrating a large diversity of data sources, resulting in 4.7M training examples from 7 tasks (NLI, QA, paraphrasing, fact verification, information retrieval, semantic similarity, summarization). In [128], the authors present the first dataset with fine-grained factual error annotations named DIASUMFACT wherein they define fine-grained factual error detection as a sentence-level multi-label classification task and evaluate two state-of-the-art models on their dataset.

In [36], the authors propose a factual knowledge benchmark called Pinocchio containing 20K diverse factual questions that span different sources, timelines, domains, geographies, and languages. Furthermore, the authors investigate whether LLMs can compose multiple facts, update factual knowledge temporally, reason over multiple pieces of facts, identify subtle factual differences, and resist adversarial examples. In [104], the authors present a benchmark, CARE-MI, for evaluating LLM misinformation in: 1) a sensitive topic, specifically the maternity and infant care domain; and 2) a language other than English, namely Chinese. In [19], the authors propose FACTOOL, a task and domain agnostic framework to detect factual errors in generated texts.

In [120], the authors construct 12K high-quality data to assess the strengths, weaknesses, and potential risks of various off-the-shelf LLMs in dealing with ambiguous information. In [12], emphasizing on both epistemological adequacy and presentational quality, the authors present a comprehensive evaluation framework, grounded in science communication research, to assess LLM responses to questions about climate change. In [66], the authors introduce Fakepedia, a counterfactual dataset designed to evaluate grounding abilities of LLMs when their internal parametric knowledge clashes with contextual information. The authors benchmark various LLMs with Fakepedia and conduct a causal

mediation analysis of LLM components when answering Fakepedia queries, based on Masked Grouped Causal Tracing (MGCT) method.

In [106], the authors investigate LLMs' susceptibility to persuasive conversations, particularly on factual questions that they can answer correctly. The authors first curate the Farm (i.e., Fact to Misinform) dataset, which contains factual questions paired with systematically generated persuasive misinformation, and develop a testing framework to track LLMs' belief changes in a persuasive dialogue. In [35], the authors introduce MAFALDA, a benchmark for fallacy classification that merges and unites previous fallacy datasets. MAFALDA comes with a taxonomy that aligns, refines, and unifies existing classifications of fallacies. In [93], the authors introduce a new dataset for FActScore on texts generated by strong multilingual LLMs. Their evaluation shows that LLMs exhibit distinct behaviors in both fact extraction and fact scoring tasks. In [90], the authors create 5Pils, a dataset of 1,676 fact-checked images with question-answer pairs about their original meta-context. Annotations are based on the five Pillars fact-checking framework.

Table 2. Deepfake Detection Solutions and Platforms

Tool	Organization	Data Type	Description
DeepStar	ZeroFox	videos	open-source research toolkit to enhance detection accuracy
Sentinel AI	Sentinel	video, audio	API based on cyber-security standard of Defense in Depth
WeVerify Deepfake detector	WeVerify	video, images	decentralized collaborative API for content verification
SynthID	Google Deepmind	audio, images, video	watermarking tool for content provenance
FakeCatcher	Intel	images, video	web platform that uses spatio-temporal maps of human markers for detecting fakes
Sensity	Sensity	audio, images, video	platform for multi-industry fake detection and ID thefts
Deepware Scanner	Deepware	videos	platform detects AI-generated manipulations of human faces
Phocus, AI text and voice detector	DuckDuckGoose	text, audio image, video	suite of 6 deepfake detectors with explanations of results
Microsoft Video AI Authenticator	Microsoft	images, videos	identifies inconsistencies in merging boundaries and subtle grayscale elements

4 Disinformation

Unlike misinformation, disinformation is intentional spread of malicious and false information. Disinformation could take several forms such as fake videos, fake news, fake reviews, fake ordering of search results/cherry picking, forgery, use of overloaded language, and so on. A list of various types of disinformation can be found in [5]. Since a majority of works in this area have focused on fake news,

we organize related work into two main topics, one related to text-based fake news and rumors, and another related to multimodal deepfakes, forgery, and fake accounts spanning image, video, and audio data.

4.1 Fake News and Rumors

Fake news and rumors are false and misleading information presented as news, targeting general public by and large. From creating chaos and confusion to fear and hate, fake news and rumors could adversely impact individuals and society. Below, we review fake news detection and mitigation methods.

Detection and Mitigation Methods. In [79], the authors propose to leverage information in the external news environment where a fake news post is created and disseminated, for detecting fake news. In [11], the authors describe a new algorithm called Preferential Attachment k-class Classifier (PreAttacK) for detecting fake accounts in a social network. Investigating the effect of adversarial attack on reviews, the authors in [20] learn a fake review generator through reinforcement learning, which maliciously promotes items by forcing prediction shifts after adding generated reviews to the system. In a related work [1], the authors study the impact of news ordering on audience perception.

In [100], the authors introduce SheepDog, a style-robust fake news detector that prioritizes content over style to the determine veracity of news. In [101], the authors propose DECOR, a novel application of Degree-Corrected Stochastic Blockmodels to the fake news detection problem. In [17], the authors propose a method based on causal intervention and counterfactual reasoning for Multimodal Fake News Detection. In [127], the authors propose a method for detecting fake news via graph neural networks. In [56], a MUlti-Step Evidence Retrieval Enhancement Framework (called MUSER) is proposed for fake news detection.

In [112], the authors propose a fake news detection model (MHDF) for multisource heterogeneous data progressive fusion. In [64], the authors assess the effectiveness of news em- beddings from ChatGPT for detecting fake news and showcase that despite their initial performance slightly surpassing the pre-trained BERT model, they still lag behind the state-of-the-art. In [117], the authors propose a method known as Diverse Counterfactual Evidence framework for Rumor Detection (DCE-RD). The authors exploit the diverse counterfactual evidence of an event graph to serve as multi-view interpretations, which are further aggregated for robust rumor detection results.

Datasets and Benchmarks. In [15], the authors construct the first LLM-Generated Misinformation Dataset called LLMFake and study how humans and machines perform in detecting fakes. In a related work [54], the authors build a comprehensive testbed by gathering texts from diverse human writings and texts generated by different LLMs towards detecting machine generated texts. In [37], the authors create a new training dataset, PropaNews, with 2,256 examples and propose a novel framework to generate training examples that are informed by

the known styles and strategies of human-authored propaganda. Specifically, the authors perform self-critical sequence training guided by natural language inference to ensure the validity of the generated articles, while also incorporating propaganda techniques, such as appeal to authority and loaded language. In [38], the authors present a thorough exploration of ChatGPT's proficiency in generating, explaining, and detecting fake news. Specifically, they employ different prompt methods to generate fake news samples, obtain features to characterize fake news based on ChatGPT's explanations, and examine ChatGPT's capacity to identify fake news.

4.2 Deepfakes and Forgery

Fake videos, audio, and images (commonly referred to as deepfakes) have become increasingly prevalent due to generative AI technologies [61], making it hard for humans to distinguish between real and fake data [69]. Deepfakes and forgery are primarily used to tarnish individual identities, but they could be used against the interest of a larger section of society as well. Researchers note that there are at least three keys in which generative AI could aid in detecting disinformation—by creating necessary training data, by providing explanations as well as by simulating user interactions [96]. Below we review recent progress in the area.

Detection and Mitigation Methods. In [8], the authors introduce a framework called Discriminative fEature dEcoupling enhanceMent (DEEM) for detecting speech forgery. In [121], the authors propose a continual learning algorithm for fake audio detection to overcome catastrophic forgetting, called Regularized Adaptive Weight Modification (RAWM). The authors in [76] provide an analysis on audio spoofing detection and discuss future prospects in the area. In [78], the authors review methods for detecting harmful memes. In [7], the authors propose a pluggable and efficient active model watermarking framework for Deepfake detection in images. In [34], the authors investigate the manifestation of potentially verifiable language cues of deception in the presence of objective truth, a distinguishing feature absent in previous text-based deception datasets. The authors show that there exists a class of detectors (algorithms) that have similar truth detection performance compared to human subjects, even when the former accesses only the language cues while the latter engages in conversations with complete access to all potential sources of cues (language and audio-visual).

Datasets and Benchmarks. Recently, the Defense Advanced Research Project Agency (DARPA) of the United States launched the 'Semantics Forensics Initiative' comprising of an analytic catalog containing open-source resources for deepfake detections and an open community research effort called AI Forensics Open Research Challenge Evaluation (AI FORCE), which aims to develop AI models that can accurately detect synthetic AI-generated images [92]. In [40], the authors present a Speech-Forensics dataset by extensively covering authentic, synthetic, and partially forged speech samples that include multiple segments

synthesized by different high-quality algorithms. The authors propose a TEmporal Speech LocalizaTion network, called TEST, aiming at simultaneously performing authenticity check, and multiple fake segments localization. In [107], the authors propose a comprehensive benchmark called DeepfakeBench for deepfake detection. In [85], a dataset for analyzing misleading video headlines is provided. In [122], a comprehensive benchmark for evaluating the safety of LLMs is provided, comprising of 11,435 diverse multiple choice questions spanning across 7 distinct categories of safety concerns. In [108], the authors present a crowd intelligence-based semantic feature learning module to capture textual content's sequential and hierarchical features and then design a knowledge-based semantic structural mining module that leverages ChatGPT for knowledge enhancement. There are also several open source tools and platforms apart from commercial ventures that are aimed at fighting deepfakes. An overview of some of these tools is provided in Table 2.

5 Human-Centric Best Practices

In addition to various detection and mitigation methods, experts from diverse fields such as AI policy, law and regulation, economics, and other scientific disciplines have proposed guidelines and suggested best practices towards addressing some of the concerns associated with AI-generated misinformation and disinformation.

Researchers note that mitigating the risks from frontier AI systems requires up-to-date and reliable information about those systems and suggest that developers should disclose safety-critical information to government actors and other developers, who can then decide on appropriate technical, organizational, and policy responses [49]. Further, it has been argued that other independent domain experts in academia and civil society should also receive key information about frontier models, and that these stakeholders should be able to provide guidance to both developers and government actors [49]. Echoing similar idea, in [29], the authors argue for closer proximity between the people involved in research and regulation.

Noting that language models are shaped by the language of those with whom they interact, in [51], the authors call upon the research community to investigate these 'societies' of interacting artificial intelligences to increase their rewards and reduce their risks for human society. The authors illustrate that decentralized AI collectives can spontaneously expand the bounds of human diversity and reduce the risk of toxic, anti-social behavior online.

Going a step beyond the above suggestions that emphasize on broadening stakeholder participation, recent works have urged for stakeholder empowerment [44]. Specifically, the authors state that there exists a power dichotomy between developers and affected stakeholders and that the current AI discourse does not address the causes that uphold such power asymmetry. The authors urge for alternate AI ownership models that govern generative AI production and distribution [44]. On a related note, researchers argue that in enabling stakeholder

empowerment, there may be a need to look beyond technical methods to fields such as the arts and the humanities [94].

6 Open Problems and Future Directions

Despite the many advancements, preventing, detecting, and mitigating misinformation and disinformation in the age of generative AI remains a significant challenge [77]. There are many open problems that ought to be addressed, some of which we list in this section.

6.1 Encapsulating Regional Contexts

Despite the vast diversity in cultures and languages spoken around the world, methods to mitigate the harms of generative AI have largely focused on popular languages such as English [6]. As a result, generative models based on low-resource languages remain far more vulnerable to misinformation and disinformation. Furthermore, there also exists a dataset gap in the sense that existing low resource language datasets do not reflect the diversity of region-specific socio-economic-cultural contexts. As a result, generative models are prone to a wide range of representational harms [83]. Therefore, datasets and models that can detect and mitigate region-specific misinformation and disinformation have to be developed.

6.2 Addressing Domain-Specific Issues

Yet another challenge concerns the development of models that can detect and mitigate domain-specific misinformation and disinformation. While domains such as finance, healthcare, and climate have received some attention [12,126], fields like the arts, sports, law, or the literature are largely unexplored. For example, popular watermarking methods do not necessarily work in all application scenarios [116]. Given the increased adoption of generative AI in these industries, it becomes imperative to develop methods that can detect domain- specific misinformation and disinformation, issues that are less likely to be prevalent in other commonly tested scenarios.

6.3 Multimodal and Real-Time Alignment

In recent years, there has been an increased focus on developing models that can leverage multimodal data to detect disinformation while also providing model explanations [57,73,86,89]. That said, these works do not necessarily address issues stemming from spatio-temporal misalignment. For example, new forms of deepfake manipulations can arise based on ongoing world events; certain socio-cultural norms may be outdated, so models need to be able to learn these spatio-temporal correlations in real time to provide robust performance.

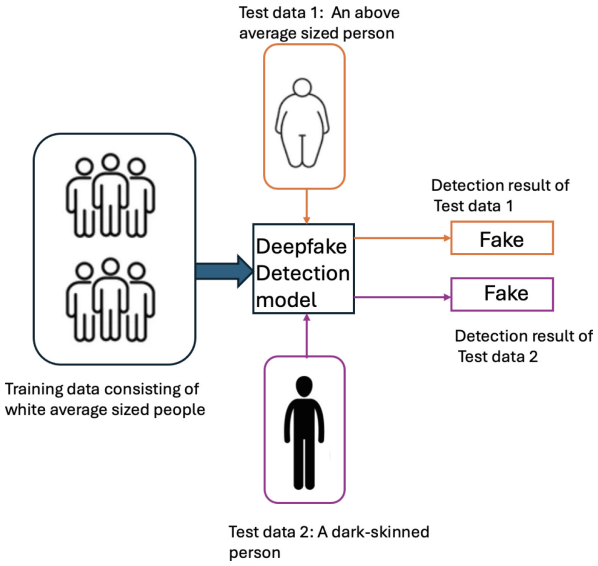

Fig. 1. A hypothetical illustration of model bias resulting in errors in deepfake detection wherein a model trained predominantly on images of white averaged sized people may classify images of fat and dark skinned people as fake although they may be real

6.4 Mitigating Biases

In a recent study, through a qualitative study of over two thousand individuals, the authors revealed that the level of vulnerability of a user to misinformation and disinformation is dependent on their own biases such as their prior knowledge, priming effect, imposter bias, homophily biases, and heterophily biases wherein their demographic background might affect the evaluation [14,62]. Studies have also shown that deepfake detection models suffer from similar biases [91], drastically affecting detection accuracy across protected attributes such as race, gender, age, etc. Please see Fig. 1 for a hypothetical illustration. This requires the development of models that optimize both fairness and accuracy across protected attributes. Research in this direction is still in its infancy [46], with more research needed to examine the generalizability of model performance, to design context-relevant fairness metrics based on local geographies, among others.

7 Conclusions

In this survey paper, we reviewed recent progress with regard to misinformation and disinformation in LLMs and more broadly generative AI models. We enlisted various countermeasures aimed towards addressing these issues by organizing the survey findings along different aspects that encompassed misinformation and disinformation detection methods, mitigation methods, datasets and benchmarks. We also provided a summary of best practices as suggested by a wider set of

stakeholders beyond the AI community. Finally, we discussed some open problems and future directions. We hope the survey helps researchers, developers, and other stakeholders in getting a comprehensive overview of the recent developments related to misinformation and disinformation in generative AI.

References

1. Advani, R., Papotti, P., Asudeh, A.: Maximizing neutrality in news ordering. In: KDD (2023)
2. Agrawal, G., Kumarage, T.S., Alghamdi, Z., liu, H.: Can knowledge graphs reduce hallucinations in LLMs?: a survey. In: NAACL (2024)
3. Akbar, S.A., et al.: Hallumeasure: fine-grained hallucination measurement using chain-of-thought reasoning. In: EMNLP (2024)
4. Aldausari, N., Sowmya, A., Marcus, N., Mohammadi, G.: Video generative adversarial networks: a review. ACM Comput. Surv. **55**(2), 1–25 (2022)
5. Alhindi, T., Chakrabarty, T., Musi, E., Muresan, S.: Multitask instruction-based prompting for fallacy recognition. In: EMNLP (2022)
6. AlKhamissi, B., et al.: Investigating cultural alignment of large language models. In: ACL (2024)
7. Bao, H., et al.: Pluggable watermarking of deepfake models for deepfake detection. In: IJCAI (2024)
8. Bei, Y., et al.: Discriminative feature decoupling enhancement for speech forgery detection. In: IJCAI (2024)
9. Bodaghi, A., et al.: A literature review on detecting, verifying, and mitigating online misinformation. IEEE Trans. Comput. Soc. Syst. (2024)
10. Bouyamourn, A.: Why LLMs hallucinate, and how to get (evidential) closure: perceptual, intensional, and extensional learning for faithful natural language generationa. In: ACL (2023)
11. Breuer, A., Khosravani, N., Tingley, M., Cottel, B.: Preemptive detection of fake accounts on social networks via multi-class preferential attachment classifiers. In: KDD (2023)
12. Bulian, J., et al.: Assessing large language models on climate information. In: ICML (2024)
13. Cao, Q., Cheng, J., Liang, X., Lin, L.: Visdiahalbench: a visual dialogue benchmark for diagnosing hallucination in large vision-language models. In: ACL (2024)
14. Casu, M., Guarnera, L., Caponnetto, P., Battiato, S.: Genai mirage: the impostor bias and the deepfake detection challenge in the era of artificial illusions. Forensic Sci. Int. Digit. Invest. **50** (2024)
15. Chen, C., Shu, K.: Can LLM-generated misinformation be detected? In: ICLR (2024)
16. Chen, C., Shu, K.: Combating misinformation in the age of LLMs: opportunities and challenges. AI Mag. (2024)
17. Chen, Z., Hu, L., Li, W., Shao, Y., Nie, L.: Causal intervention and counterfactual reasoning for multi-modal fake news detection. In: ACL (2023)
18. Cheng, X., et al.: Small agent can also rock! empowering small language models as hallucination detector. In: EMNLP (2024)
19. Chern, I.C., et al.: Factool: factuality detection in generative AI a tool augmented framework for multi-task and multi-domain scenarios. ArXiv (2023)

20. Chiang, H.Y., et al.: Shilling black-box review-based recommender systems through fake review generation. In: KDD (2023)
21. Chu, Z., et al.: Sora detector: a unified hallucination detection for large text-to-video models. ArXiv (2024)
22. Du, Y., Li, S., Torralba, A., Tenenbaum, J., Mordatch, I.: Assessing large language models on climate information. In: ICML (2024)
23. Eiras, F., et al.: Position: near to mid-term risks and opportunities of open-source generative AI. In: ICML (2024)
24. Elazar, Y., et al.: What's in my big data? In: ICLR (2024)
25. Feffer, M., Sinha, A., Deng, W.H., Lipton, Z.C., Heidari, H.: Red-teaming for generative AI: silver bullet or security theater? In: AIES (2024)
26. Gallegos, I.O., et al.: Bias and fairness in large language models: a survey. Comput. Linguist. **50**(3) (2024)
27. Geigle, G., Timofte, R., Glavaš, G.: Does object grounding really reduce hallucination of large vision-language models? In: EMNLP (2024)
28. Gekhman, Z., Herzig, J., Aharoni, R., Elkind, C., Szpektor, I.: Trueteacher: learning factual consistency evaluation with large language models. In: EMNLP (2023)
29. Goanta, C., Aletras, N., Chalkidis, I., Ranchordás, S., Spanakis, G.: Regulation and NLP (regNLP): taming large language models. In: EMNLP (2023)
30. Golchin, S., Surdeanu, M.: Time travel in LLMs: tracing data contamination in large language models. In: ICLR (2024)
31. Guan, J., Dodge, J., Wadden, D., Huang, M., Peng, H.: Language models hallucinate, but may excel at fact verification. In: NAACL (2024)
32. Halawi, D., Denain, J.S., Steinhardt, J.: Overthinking the truth: understanding how language models process false demonstrations. In: ICLR (2024)
33. Han, T., et al.: The instinctive bias: spurious images lead to hallucination in MLLMs. In: EMNLP (2024)
34. Hazra, S., Majumder, B.P.: To tell the truth: language of deception and language models. In: NAACL (2024)
35. Helwe, C., Calamai, T., Paris, P.H., Clavel, C., Suchanek, F.M.: Mafalda: a benchmark and comprehensive study of fallacy detection and classification. In: NAACL (2024)
36. Hu, X., et al.: Towards understanding factual knowledge of large language models. In: ICLR (2024)
37. 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: ACL (2023)
38. Huang, Y., Sun, L.: FakeGPT: fake news generation, explanation and detection of large language model. ArXiv (2024)
39. Jaidka, K., et al.: Misinformation, disinformation, and generative AI: implications for perception and policy. Res. Pract. Digit. Gov. (2024)
40. Ji, Z., Lin, C., Wang, H., Shen, C.: Speech-forensics: towards comprehensive synthetic speech dataset establishment and analysis. In: IJCAI (2024)
41. Ji, Z., Gu, Y., Zhang, W., Lyu, C., Lin, D., Chen, K.: Anah: analytical annotation of hallucinations in large language models. In: ACL (2024)
42. Ji, Z., et al.: Survey of hallucinations in natural language generation. ACM Comput. Surv. (2023)
43. Jiang, C., et al.: On large language models' hallucination with regard to known facts. In: NAACL (2024)
44. Jonne Maas, A.M.I.: Beyond participatory AI. In: AIES (2024)
45. Ju, Y., et al.: Deepfake-o-meter v2.0: an open platform for deepfake detection. ArXiv (2024)

46. Ju, Y., et al.: Improving fairness in deepfake detection. In: WACV (2024)
47. Kapoor, S., et al.: Position: On the societal impact of open foundation models. In: ICML (2024)
48. Kawakami, A., Wilkinson, D., Chouldechova, A.: Do responsible AI artifacts advance stakeholder goals? four key barriers perceived by legal and civil stakeholders. In: AIES (2024)
49. Kolt, N., et al.: Responsible reporting for frontier AI development. In: AIES (2024)
50. Lai, J., et al.: RumorLLM: a rumor large language model-based fake-news-detection data-augmentation approach. Appl. Sci. (2024)
51. Lai, S., Potter, Y., Kim, J., Zhuang, R., Song, D., Evans, J.: Position: evolving AI collectives enhance human diversity and enable self-regulation. In: ICML (2024)
52. Le, M., et al.: Voicebox: text-guided multilingual universal speech generation at scale. ArXiv (2023)
53. Li, J., et al.: The dawn after the dark: an empirical study on factuality hallucination in large language models. In: ACL (2024)
54. Li, Y., et al.: Mage: machine-generated text detection in the wild. In: ACL (2024)
55. Li, Y., Wang, S., Ding, H., Chen, H.: Large language models in finance: a survey. In: ICAIF (2023)
56. Liao, H., et al.: Muser: a multi-step evidence retrieval enhancement framework for fake news detection. In: KDD (2023)
57. Liu, H., Wang, W., Li, H., Li, H.: Teller: a trustworthy framework for explainable, generalizable and controllable fake news detection. In: ACL (2024)
58. Liu, T., et al.: A token-level reference-free hallucination detection benchmark for free-form text generation. In: ACL (2022)
59. Liu, Y., et al.: Decoding susceptibility: modeling misbelief to misinformation through a computational approach. In: EMNLP (2024)
60. Liu, Y., Ji, T., Sun, C., Wu, Y., Zhou, A.: Investigating and mitigating object hallucinations in pretrained vision-language (clip) models. In: EMNLP (2024)
61. Loth, A., Kappes, M., Pahl, M.O.: Blessing or curse? a survey on the impact of generative AI on fake news. ArXiv (2024)
62. Lovato, J., et al.: Diverse misinformation: impacts of human biases on detection of deepfakes on networks. Nature (2024)
63. Lucas, J., Uchendu, A., Yamashita, M., Lee, J., Rohatgi, S., Lee, D.: Fighting fire with fire: the dual role of LLMs in crafting and detecting elusive disinformation. In: EMNLP (2023)
64. Ma, X., Zhang, Y., Ding, K., Yang, J., Wu, J., Fan, H.: On fake news detection with LLM enhanced semantics mining. In: EMNLP (2024)
65. Mendes, E., Chen, Y., Xu, W., Ritter, A.: Human-in-the-loop evaluation for early misinformation detection: a case study of COVID-19 treatments. In: ACL (2023)
66. Monea, G., et al.: A glitch in the matrix? locating and detecting language model grounding with fakepedia. In: ACL (2024)
67. Mueller, F.B., Görge, R., Bernzen, A.K., Pirk, J.C., Poretschkin, M.: LLMs and memorization: On quality and specificity of copyright compliance. In: AIES (2024)
68. Nan, Q., Sheng, Q., Cao, J., Hu, B., Wang, D., Li, J.: Let silence speak: enhancing fake news detection with generated comments from large language models. In: CIKM (2024)
69. Nguyen, T.T., et al.: Deep learning for deepfakes creation and detection: a survey. Comput. Vis. Image Underst. (2022)
70. Niu, C., et al.: Ragtruth: a hallucination corpus for developing trustworthy retrieval-augmented language models. In: ACL (2024)

71. Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: interactive simulacra of human behavior. In: UIST (2023)
72. Petryk, S., et al.: Aloha: a new measure for hallucination in captioning models. In: NAACL (2024)
73. Qi, P., Yan, Z., Hsu, W., Lee, M.L.: Sniffer: multimodal large language model for explainable out-of-context misinformation detection. In: CVPR (2024)
74. Qiu, Y., Ziser, Y., Korhonen, A., Ponti, E., Cohen, S.: Detecting and mitigating hallucinations in multilingual summarisation. In: EMNLP (2023)
75. Ramprasad, S., Ferracane, E., Lipton, Z.C.: Analyzing LLM behavior in dialogue summarization: unveiling circumstantial hallucination trends. In: ACL (2024)
76. Ranjan, R., Vatsa, M., Singh, R.: Uncovering the deceptions: An analysis on audio spoofing detection and future prospects. In: IJCAI (2023)
77. Rauh, M., et al.: Gaps in the safety evaluation of generative AI. In: AIES (2024)
78. Sharma, S., et al.: Detecting and understanding harmful memes: a survey. In: IJCAI-ECAI (2022)
79. Sheng, Q., Cao, J., Zhang, X., Li, R., Wang, D., Zhu, Y.: Zoom out and observe: news environment perception for fake news detection. In: ACL (2022)
80. Shi, W., Han, X., Lewis, M., Tsvetkov, Y., Zettlemoyer, L., Tau Yih, W.: Trusting your evidence: hallucinate less with context-aware decoding. In: NAACL (2024)
81. Si, C., Goyal, N., Wu, T., Zhao, C., Feng, S., III, H.D., Boyd-Graber, J.L.: Large language models help humans verify truthfulness – except when they are convincingly wrong. In: NAACL (2024)
82. Snyder, B., Moisescu, M., Zafar, M.B.: On early detection of hallucinations in factual question answering. In: KDD (2024)
83. Solaiman, I., et al.: Evaluating the Social Impact of Generative AI Systems in Systems and Society. Oxford University Press, Oxford Handbook on the Foundations and Regulation of Generative AI (2023)
84. Sui, P., Duede, E., Wu, S., So, R.: Confabulation: the surprising value of large language model hallucinations. In: ACL (2024)
85. Sung, Y., Boyd-Graber, J., Hassan, N.: Not all fake news is written: a dataset and analysis of misleading video headlines. In: EMNLP (2023)
86. Tahmasebi, S., Müller-Budack, E., Ewerth, R.: Multimodal misinformation detection using large vision-language models. In: CIKM (2024)
87. Tang, L., et al.: Tofueval: evaluating hallucinations of LLMs on topic-focused dialogue summarization. In: NAACL (2024)
88. Taveekitworachai, P., Abdullah, F., Thawonmas, R.: Null-shot prompting: rethinking prompting large language models with hallucination. In: EMNLP (2024)
89. Tian, J.J., et al.: Web retrieval agents for evidence-based misinformation detection. In: COLM (2024)
90. Tonglet, J., Moens, M.F., Gurevych, I.: Image, tell me your story!" predicting the original meta-context of visual misinformation. In: EMNLP (2024)
91. Trinh, L., Liu, Y.: An examination of fairness of AI models for deepfake detection. In: IJCAI (2021)
92. Trinh, L., Liu, Y.: Semantic forensics (2024). https://www.darpa.mil/program/semantic-forensics
93. Vu, K.T., Krumdick, M., Reddy, V., Dernoncourt, F., Lai, V.D.: An analysis of multilingual factscore. In: EMNLP (2024)
94. Walker, J., Thuermer, G., Vicens, J., Simperl, E.: AI art and misinformation: approaches and strategies for media literacy and fact checking. In: AIES (2023)

95. Wan, F., Huang, X., Cui, L., Quan, X., Bi, W., Shi, S.: Knowledge verification to nip hallucination in the bud. In: EMNLP (2024)
96. Wan, H., Feng, S., Tan, Z., Wang, H., Tsvetkov, Y., Luo, M.: Dell: generating reactions and explanations for LLM-based misinformation detection. In: ACL (2024)
97. Wang, W., Haddow, B., Birch, A., Peng, W.: Assessing factual reliability of large language model knowledge. In: NAACL (2024)
98. Whitney, C.D., Norman, J.: Real risks of fake data: synthetic data, diversity-washing and consent circumvention. In: FAccT (2024)
99. Widder, D.G., Nafus, D., Dabbish, L., Herbsleb, J.: Limits and possibilities for "ethical AI" in open source: a study of deepfakes. In: FAccT (2022)
100. Wu, J., Guo, J., Hooi, B.: Fake news in sheep's clothing: robust fake news detection against LLM-empowered style attacks. In: KDD (2024)
101. Wu, J., Hooi, B.: Decor: degree-corrected social graph refinement for fake news detection. In: KDD (2023)
102. Wu, J., Chao, W., Zhou, X., Luo, Z.: Characterizing and verifying scientific claims: qualitative causal structure is all you need. In: EMNLP (2023)
103. Xia, Y., et al.: Hallucination diversity-aware active learning for text summarization. In: NAACL (2024)
104. Xiang, T., et al.: Care-mi: Chinese benchmark for misinformation evaluation in maternity and infant care. In: NeurIPS Datasets and Benchmarks Track (2023)
105. Xu, H., Gan, W., Qi, Z., Wu, J., Yu, P.S.: Large language models for education: a survey. J. Mach. Learn. Cybern. (2024)
106. Xu, R., et al.: The earth is flat because...: investigating LLMs' belief towards misinformation via persuasive conversation. In: ACL (2024)
107. Yan, Z., Zhang, Y., Yuan, X., Lyu, S., Wu, B.: Deepfakebench: a comprehensive benchmark of deepfake detection. In: NeurIPS Datasets and Benchmarks Track (2023)
108. Yang, C., Zhang, P., Qiao, W., Gao, H., Zhao, J.: Rumor detection on social media with crowd intelligence and ChatGPT-assisted networks. In: EMNLP (2023)
109. Yao, F., Li, C., Nekipelov, D., Wang, H., Xu, H.: Human vs. generative AI in content creation competition: symbiosis or conflict? In: ICML (2024)
110. Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., Zhang, Y.: A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. High Confidence Comput. (2024)
111. Yehuda, Y., Malkiel, I., Barkan, O., Weill, J., Ronen, R., Koenigstein, N.: InterrogateLLM: zero-resource hallucination detection in LLM-generated answers. In: ACL (2024)
112. Yu, Y., Ji, K., Gao, Y., Chen, Z., Ma, K., Wu, J.: MHDF: multi-source heterogeneous data progressive fusion for fake news detection. In: PAKDD (2024)
113. Yuan, H., et al.: Whispers that shake foundations: analyzing and mitigating false premise hallucinations in large language models. In: EMNLP (2024)
114. Yue, Z., Zhang, L., Jin, Q.: Less is more: mitigating multimodal hallucination from an EOS decision perspective. In: ACL (2024)
115. Zha, Y., Yang, Y., Li, R., Hu, Z.: Alignscore: evaluating factual consistency with a unified alignment function. In: ACL (2023)
116. Zhang, H., Edelman, B., Francati, D., Venturi, D., Ateniese, G., Barak, B.: Watermarks in the sand: impossibility of strong watermarking for language models. In: ICML (2024)
117. Zhang, K., et al.: Rumor detection with diverse counterfactual evidence. In: KDD (2023)

118. Zhang, M., Choi, E.: Mitigating temporal misalignment by discarding outdated facts. In: EMNLP (2023)
119. Zhang, M., Press, O., Merrill, W., Liu, A., Smith, N.: How language model hallucinations can snowball. In: ICML (2024)
120. Zhang, T., et al.: Clamber: a benchmark of identifying and clarifying ambiguous information needs in large language models. In: ACL (2024)
121. Zhang, X., Yi, J., Tao, J., Wang, C., Zhang, C.Y.: Do you remember? overcoming catastrophic forgetting for fake audio detection. In: ICML (2023)
122. Zhang, Z., et al.: Safetybench: evaluating the safety of large language models. In: ACL (2024)
123. Zhao, W.X., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2023)
124. Zhong, M., Chen, K., Xue, Z., Liu, L., Yang, M., Zhang, M.: On the hallucination in simultaneous machine translation. In: ACL (2024)
125. Zhong, W., et al.: Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In: ACL (2024)
126. Zhu, L., Mou, W., Luo, P.: Potential of large language models as tools against medical disinformation. JAMA Internal Med. (2024)
127. Zhu, P., Pan, Z., Liu, Y., Tian, J., Tang, K., Wang, Z.: A general black-box adversarial attack on graph-based fake news detectors. In: IJCAI (2024)
128. Zhu, R., Qi, J., Lau, J.H.: Annotating and detecting fine-grained factual errors for dialogue summarization. In: ACL (2023)