SHORT-PAPER

Online Disinformation and Generative Language Models: Motivations, Challenges, and Mitigations

ZIYI GUO, Trinity College Dublin, Dublin, Leinster, Ireland

Published: 13 May 2024

Citation in BibTeX format

WWW '24: The ACM Web Conference 2024
May 13 - 17, 2024
Singapore, Singapore

Conference Sponsors:
SIGWEB

WWW '24: Companion Proceedings of the ACM Web Conference 2024 (May 2024)
https://doi.org/10.1145/3589335.3651254
ISBN: 9798400701726

Online Disinformation and Generative Language Models:

Motivations, Challenges, and Mitigations

Ziyi Guo
Supervised by Owen Conlan
ADAPT Centre, Trinity College Dublin
Dublin, Ireland
guozi@tcd.ie

ABSTRACT

Disinformation refers to the deliberate dissemination of fake or misleading information, which significantly threatens the modern social stability by undermining trust, intensifying polarization and manipulating public opinion. With the advances of generative AI, the landscape of modern disinformation is changing following the rise of Large Language Models. Recent studies has revealed the capability of generative language models to create convincing and misleading content against the truth and warned the availability of such models to be maliciously abused for deceptive generation.

However, AI-driven disinformation is a human-centered societal issue in nature, the realization of which requires not only the in-depth discussion on the latest trends from both sides of generative AI and disinformation, but a critical analysis on the uncertainty of their potential interaction in practice as well. The paper introduces the new vision of AI-driven disinformation campaigns from the perspectives of human-centered AI, proposes a framework of core research questions based on the existing research gap, discusses the preliminary discovery in literature and initial experiments, and elaborates the main lines of research in the future work.

CCS CONCEPTS

• **Human-centered computing** → Human computer interaction.

KEYWORDS

Disinformation, Large Generative Language Models, Motivations, Generations, Mitigations

ACM Reference format:

Ziyi Guo. 2024. Online Disinformation and Generative Language Models: Motivations, Challenges, and Mitigations. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13-17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi. org/10.11 45/3589335.3651254

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05.
https://doi.org/10.1145/3589335.3651254

1 INTRODUCTION

Following the 2016 US Presidential Election, the concept of *fake news* has gained significant attention in the academia and media world. The term shares the commonalities in nature with the long-standing phenomenon *disinformation* throughout the history [1], which is spread as the *covert* or *deceptive* efforts to influence the opinions of a target audience [2, 3]. Beyond the early paradigms of disinformation campaigns based on traditional media [4], the advancing internet has fundamentally changed the way of (dis-) information sharing. Such transformation heralds an increase in the volume, efficiency, and scope of disinformation, intensifying the challenge for individuals to discern between the authentic and fabricated information, while also fostering social unrest [5].

In recent years, the advancement of generative artificial intelligen-ce (GAI) technology has gained increasing attention even beyond the computer science community [6]. As the latest breakthrough in the field of GAI, Large Language Models (LLMs) represented by ChatGPT demonstrate tremendous power in natural language understanding and generation and can contribute to a considerable number of fields [7, 8]. However, concerns have been raised on the ethical issues of the models at the same time [9], regarding their misuse as the potential promotion factors for disinformation creation and dissemination. With the capacity of LLM, actors with malicious purpose can produce convincing and context-sensitive content effortlessly [10], which could be easily weaponized to fuel disinformation campaigns [11].

Therefore, this work is established on the foundations of existing research around the characterization of disinformation campaigns, the capacity of generative language models associated with ethical issues, and their potential correlations embodied as LLM-driven disinformation campaigns. Considering the shifting landscape of both disinformation and LLM research, this work focuses on the identification of core elements of disinformation, demonstrates the emerging challenge driven by LLM on effective disinformation campaigns with theoretical and empirical evidences and rationally analyzes the gaps in existing research concerning the practical applications of the proposed challenge. Potential mitigations in response are discussed as well. In all, This work is expected to contribute to the future development of disinformation research in the era of generative language models.

Phase	Generative Model & Human Factor-based Challenges		AI & Human-centered Mitigations
Disinformation Generation	Generative Language Model	Automation	Training Data Harmlessification
		Persuasiveness Enhancement	Model Safety Enhancement
		Camouflage Enhancement	AI-generated Content Detection
	Human Factor	Increasing Actors	Access Control
		Rising Disinformation Scale	Ethical Norms Development
		Speculation	Usage Restriction
Disinformation Dissemination	Generative Language Model	Personalization	Collaborative Behavior Detection
		Dynamization	Personalized Counterfactual Norms
		Contexualization	Radiocative Content Marker
	Human Factor	More Diverse Groups	Humanhood Verification
		Efficient-Behavioral Campaign	Social Early Intervention
		Novel Tactics	Media Literacy Campaign

Table 1: An Overview of *LLMs Meet Disinformation*

2 RESEARCH PROBLEM

With the contextual identification of the serious challenge induced by emerging technologies, a core research framework with a set of related problems with clarifications is hereby proposed as follows:

RQ 1. What can generative language models do to facilitate the effectiveness of disinformation generation? Recent progress in generative AI techniques demonstrates the enormous potential of language models to rival human-written content at relatively low costs. However, disinformation is a long-lasting societal issue of developing demonstrations, thus challenging the effectiveness of such AI-generated disinformation in terms of persuasiveness and camouflage in times. Besides, choosing between the generic foundation LLM and the specific fine-tuned lightweight model for disinformation generation is also an outstanding issue considering the unexplored effectiveness and consumption of both methods in the space. Such trade-off issue requires further verification.

RQ 2. Why use generative language models to promote online disinformation dissemination? Such proposal requires sufficient justification. Despite of the raising concerns regarding the misuse of LLMs for harmful generation technically, online disinformation is human-centered movement in nature based on the actor intent. Whether baseline models can fuel deceptive tactics in an assumed *Actor-LLM-Disinformation* framework, and whether the existing models with ever-enhanced ethical and safety regulations produce disinformation of higher quality at lower cost than manual work remains in doubt in practice. Considerations on the human factors in the scenario should be given regarding actor motivations, based on the application of comprehensive evaluation on the usability, reliability and efficiency of model as a human-center AI system.

RQ 3. How to mitigate the online disinformation campaign in the era of generative language models? Undergoing the changes with the impact of emerging technology, the online disinformation campaign with social implications potentially enhanced by the AI-generated content becomes the objective against the traditional mitigations. Discussions on all levels from generation to dissemination need to be conducted concerning not only the effectiveness of current methods, but the potential enhancement driven by large language models as well. In the context of a complicated societal issue, integrating algorithmic and human-engaged evaluations is crucial for building effective and robust mitigations.

3 RELATED WORK

3.1 Disinformation and Influence Operations

Disinformation is the deliberate spread of false and misleading content against truth, deceptively targeting the audiences' opinion in influence operations [1, 2]. The acknowledged *ABC* framework models disinformation with *actor*, *behavior* and *content* as its core facets [12]. Studies elaborate on such model through goals, which impact domestic culture, governance and policy, and manipulate nation images in international diplomacy [13]. Hiding the truth from audience, domestic actors are more worrisome than foreign power [14]. In measuring impact, studies encounter challenges in quantifying persuasion or distraction at content level [15, 16], resorting to feature-based analysis concerning traffic resource [17], content quality [18] and detectability [19], additionally with the erosion of societal trust as the long-term consequence [20].

3.2 Generative Language Models

Generative language models refers to the AI system learning from extensive amount of textual data to understand the patterns and generate new data [8]. State-of-the-art representatives are the *GPT*, *Llama* and *Mixtral* families [22, 23, 24]. Driven by the advances of data scale, foundation models and computing powers, the large language models can perform human-level text understanding and production aligned with the human preference [8]. Despite of the strong capacities, such models still suffer from the deficiencies of length-limited input context and generation [7] and also untruthful, hallucinated and unethical generations [9].

3.3 Generative Language Models and Disinformation

Generative AI has its history with disinformation, known as the Deep Fake problem [25]. The emergence of large language model re-raises the issue at a stage where textual space is yet unexplored, with the generations and models of high quality and accessibility.

On the *actor* dimension, early studies prove the feasibility of fully automated post generation based on specialized fine-tuning model with public availability [26] and the camouflage of such synthetic generation [27], which enable the wider engagement of the *actors* and related stakeholders with political and financial aims [28]. On the *behavior* dimension, current research states the model effects in the cost reduction and scale expansion of existing behaviors by replacing human in cross-channel testing and falsification [29, 30], and also the promotion of unknown behaviors with novel tactics such as personalization and dynamization modelling [31, 32]. And on the *content* dimension, language models are capable for both commentary short-form and long-form disinformation generation with higher persuasiveness, lower detectability and repeatability comparing to the human content [10, 33, 34, 35].

4 RESEARCH APPROACH AND METHODOLOGY

In the presence of a large amount of existing works in the venue, the research objectives listed in Section 2 are firmly supported by the literature while focusing on the questions not yet answered. As a research in the field of human-computer interaction in nature, empirical, design and experimental approaches will be applied to understand the disinformation campaigns, predict the impact of potential threats and evaluate the effectiveness of the techniques.

For the generation phase aiming at model capacity, human data will be collected from existing disinformation dataset, e.g. [33] and [10] and synthetic data will be collected from the state-of-the-art large language models with the prompt engineering methods in literature [10, 35] and any other methods shown effective in the course of experiments. Despite of the unknown outcome, more perfect AI-generated content is generally expected before fine-grained evaluation, for which massive attempts will be made on the various possible prompting methods. For the dissemination phase aiming at operation effectiveness, experiment will be based on simulation approach where an online social network consisting of language model-based agents which have been prompted to somehow response to others is simulated. *Actor* is engaged in the network as the critical node of disinformation provenance. The realizability of such LLM-based dissemination network and its impact are expected to be verified.

For the works in the phases above, a commonality is the essence of testing the model usability, reliability and effectiveness from the perspectives of the actors, while the distinctions are that the former requires evaluations on content-focused response and the latter relies on integrated assessment in a connected network. For the former, core evaluation metrics will be the content quality and persuasiveness measured by textual features and human-engaged reflections [10, 33]. Besides, a novel metric of *prompt complexity* will be used to measure the extent of efforts as a crucial criteria of model usability. As for the latter, generations in the network will be taken as a whole for textual analysis. The network features will also be applied to describe the operation process [36]. Human-engaged experiment will be used to measure the impact as well, which indicates the model capacities as a result.

For the mitigation phase, experimental methods with algorithmic solutions are the main focus, including the effective detection of existing and AI-generated disinformation content and campaigns. Algorithmic evaluations on the potential algorithms will be based on the general metrics of detection and network analysis tasks [35, 36]. Human-engaged evaluation with designed questionnaires will be applied when necessary to measure the human-based impacts. Another consideration is to design mitigation based on contextual inquiry and theoretical methods, such as proposing the guideline for modern media literacy campaigns for public or the AI safety regulations for developers. The necessity and details of such is yet continued to be discussed according to the experimental results.

5 PROGRESS UP TO DATE

Ziyi Guo is a first-year PhD student undertaking a four-year full-time PhD program in computer science under the supervision of Professor Owen Conlan at Trinity College Dublin. The expected time of graduation is early 2027. During the first academic year, works have been focused on defining the research area and the research question. An extensive survey of related work has thus been undertaken and an opinion paper around the potentials of language models in multi-media disinformation campaigns has been produced as a result. Meanwhile, from the explosive amount of research regarding the social implications of language models, significant gap of existing research is observed, which inspires the formation of the doctoral research framework.

In the course of the preliminary experiment based on the literature, extensive works aiming at the validations of existing findings are carried out through the language model interfaces provided by the builders. A significant discovery presented is that the current large language models are undergoing ever-enhanced safety regulations and ethical control as time goes by. Up to the date of this paper, a considerable amount of existing works related to the LLM-driven misinformation generation such as [10], [33] and [35] have been already irreproducible based on the details from the researchers, indicating the rapid shifting landscape of the model development. Despite of the high possibility that novel malicious tactics of misusing such models would be figured out in the future, it is of significance to critically analyze the comprehensive availability of large language models in promoting large-scale disinformation campaigns in the future.

6 CONCLUSION AND FUTURE WORK

In this work, the societal issue of online disinformation in the era of generative language models is comprehensively discussed in terms of the motivations, challenges and mitigations. The current conclusions are as follows. Firstly, generative language models are impacting the future of online disinformation campaigns by enhancing content, promoting behaviors and engaging actors, but the rapid development calls the usability, reliability and efficiency of the model into question in this scenario which makes the actors' motivations seem unclear. Secondly, the measurement of online disinformation impact has always been an outstanding issue, es-pecially in the coming age of generative AI. It is thus crucial to conduct adequate algorithmic and human-engaged experiments followed by evaluations with rational criteria, in order to provide subsequent research in the field with valuable references. Finally, generative language models indeed induce new challenges in the domain of disinformation while bringing about new opportunities to play as the countering expert as well. Fully understanding the disinformation in history and the threat of AI-based promotion is needed for establishing targeted mitigations and discussing the feasibility of such in practice.

Future work is undergoing development within the framework of the proposed research. Firstly, comprehensive evaluations will be carried out on the current capacity of large language models in disinformation generation. Lightweight *dark* language model with task-oriented fine-tuning will be explored if performable. Then, attempt will be made to simulate the disinformation dissemination with the synthetic social network consisting of LLM-bots, in order to measure the impact in practice. Finally, mitigations from both content and societal will be explored according to the results of the implemented evaluation works.

ACKNOWLEDGMENTS

This paper is conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence (CRT AI) under Grant No. 18/CRT/6223. The work is supported by the ADAPT Centre funded by Science Foundation Ireland Research Centers Programme and co-funded by the European Regional Development Fund (Grant No. 13/RC/2106).

REFERENCES

[1] Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094. https://doi.org/10.1126/ science.aao2998.

[2] Goldstein, J. (2021). Foreign influence operations in the cyber age [PhD thesis]. University of Oxford. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.840171.

[3] Nimmo, B. (2020). The breakout scale: Measuring the impact of influence operations. *Brookings Foreign Policy.* https://www.brookings.edu/articles/the-break-out-scale-measuring-the-impact-of-influence-operations/.

[4] Edward, M. (1925, October 1). Fake news and the public [Article]. *Harper's Magazine.* https://harpers.org/archive/1925/10/fake-news-and-the-public/

[5] World Economic Forum (2012). Global Risks 2012 Seventh Edition. https://www.we forum.org/publications/global-risks-2012-seventh-edition/.

[6] Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226.* http://arxiv.org/abs/2303.04226/.

[7] Zhang, C., Zhang, C., Li, C., Qiao, Y., Zheng, S., Dam, S. K., ... & Hong, C. S. (2023). One small step for generative ai, one giant leap for agi: A complete survey on chatgpt in aigc era. *arXiv preprint:2304.06488.* http://arxiv.org/abs/2304.06488/.

[8] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223.* http://a rxiv.org/abs/2303.18223/.

[9] Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867.* http://arxiv.org/abs/ 2301.12867/.

[10] Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis) informs us better than humans. *arXiv:2301.11924.* http://arxiv.org/abs/2301.11924/.

[11] Buchanan, B., Lohn, A., & Musser, M. (2021). *Truth, lies, and automation: How language models could change disinformation.* Center for Security and Emerg-ing Technology. https://doi.org/10.51593/2021CA003/.

[12] François, C. (2020). Actors, behaviors, content: A disinformation ABC. *Algori-thms.* https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/05/ABC _Framework _TWG_Francois_Sept_2019.pdf/.

[13] DiResta, R., Goldstein, J. A., & Grossman, S. (2021). Middle East Influence Operations: Observations Across Social Media Takedowns. *Digital Activism and Au-thoritarian Adaptation in the Middle East*, 91. https://pomeps.org/middle-east-influ-ence-operations-observations-across-social-mediatakedowns/.

[14] Emerson T. B., &Jacob, S. (2020). Americans Were Worried About the Wrong Threat, *Atlantic.* https://www.theatlantic.com/ideas/archive/2021/01/bigger-threat-wa s-always-domestic/617618/.

[15] King, G., Pan, J., & Roberts, M. E. (2017). How the Chinese government fabri-cates social media posts for strategic distraction, not engaged argument. *American political science review*, *111*(3), 484-501. https://doi.org/10. 1017/S00030554170001 44/.

[16] Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., ... & Volfovsky, A. (2020). Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. *Proceedings of the national academy of sciences*, *117*(1), 243-250. https://doi. org/10.1073/pnas.1906420116/.

[17] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. *Science*, *363*(6425), 374-378. https://doi.org/10.1126/science.aau2706/.

[18] Park, H. S., Levine, T. R., Kingsley Westerman, C. Y., Orfgen, T., & Foregger, S. (2007). The effects of argument quality and involvement type on attitude formation and attitude change: A test of dual-process and social judgment predictions. *Human Communication Research*, *33*(1), 81-102. https://doi.org/10.1111 /J.1468-2958.2007.00290.X/.

[19] DiResta, R., & Goldstein, J. A. (2021). China's fake twitter accounts are tweet-ing into the void. *Foreign Policy.* https://foreignpolicy.com/2021/12/15/china- twitt-er-trolls-ccp-influence-operations-astroturfing/.

[20] Mazarr, M. J., Bauer, R. M., Casey, A., Heintz, S., & Matthews, L. J. (2019). The emerging risk of virtual societal warfare. *Social manipulation in a changing information environment. Santa Monica CA: RAND Corporation.* https://doi.org/10. 7249/RR2714/.

[21] Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, *107*, 1753. https://doi.org/10.15779/ Z38RV0D15J/.

[22] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774.* https://doi.org/10.48550/arXiv.2303.08774/.

[23] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288.* https://doi.org/10.48550/arXiv.2307.09288/.

[24] Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., , ... & Sayed, W. E. (2024). Mixtral of experts. *arXiv :2401.04088.* https:// doi.org/10.48550/arXiv.2401.04088/.

[25] Zhang, T. (2022). Deepfake generation and detection, a survey. Multimedia Tools and Applications, 81(5), 6259-6276. https://doi.org/10.1007/s11042-021-11733-y/.

[26] Kurenkov, A. (2022). Lessons from the GPT-4Chan Controversy. *The Gradient.* https://thegradient.pub/gpt4chan- lessons/.

[27] Will, K. (2021). AI-Powered Text From This Program Could Fool the Government. *Wired.* https://www.wired.com/story/ai-powered-text-program-could-fool-government/.

[28] Sedova, K. (2021). *AI and the Future of Disinformation Campaigns: Part 2, a Threat Model.* Center for Security and Emerging Technology. https://cset.georgetown. edu/wp-content/uploads/CSET-AI-and-the-Future-of-Disinformation-Campaigns-Part-2.pdf

[29] Senate Report No 116-290, vol 2. (2020). Russian Active Measures Campaigns and Interference in the 2016 U.S. Election https://www.intelligence.senate.gov/sites /default/files/documents/Report_ Volume2.pdf.

[30] Hitlin, P., Olmstead, K., & Toor, S. (2017). Public comments to the federal communications commission about net neutrality contain many inaccuracies and duplicates.https://policycommons.net/artifacts/617566/public-comments-to-the-feder al-communications-commission-about-net-neutrality-contain-many-inaccuracies-and -duplicates/1598 389/ on 05 Feb 2024. CID: 20.500.12592/5490k9.

[31] Lohn, A., & Jackson, K. (2022). Will AI make cyber swords or shields?. https:// doi.org/10.51593/2022CA002/.

[32] Altay, S., Hacquin, A. S., Chevallier, C., & Mercier, H. (2023). Information delivered by a chatbot has a positive impact on COVID-19 vaccines attitudes and intentions. *Journal of Experimental Psychology: Applied*, *29*(1), 52. https://doi.org/1 0.1037/xap0000400/.

[33] Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023). Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–20. https://doi.org/10.1145/3544548.3581 318/.

[34] Goldstein, J. A., Chao, J., Grossman, S., Stamos, A., & Tomz, M. (2023). Can AI Write Persuasive Propaganda?.https://osf.io/fp87b/download

[35] Chen, C., & Shu, K. (2023). *Can LLM-Generated Misinformation Be Detected?* (arXiv:2309.13788). arXiv. http://arxiv.org/abs/2309.13788/.

[36] Knoke, D., & Yang, S. (2019). *Social network analysis.* SAGE publications. https://books.google.ie/books?hl=en&lr=&id=9jDEDwAAQBAJ&oi=fnd&pg=PP12&dq=n etwork+analysis&ots=HnQ4JQbn-e&sig=wn4mSp9CMOAa-gYmhyE0yNCf Iec&redir_esc=y#v=onepage&q=network%20analysis&f=false/.