Fragile, Handle With Care: On the Responsible Use of LLMs for Disinformation Detection in Human-Centered Content Moderation

Barbarella Petz
HU Berlin
Berlin, Germany
barbarella.petz@hu-berlin.de

Christopher Katins
HU Berlin
Berlin, Germany
christopher.katins@hu-berlin.de

Thomas Kosch
HU Berlin
Berlin, Germany
thomas.kosch@hu-berlin.de

Abstract

Disinformation poses a growing challenge to content moderation on online platforms, motivating the development of automated approaches for information verification, such as Large Language Model (LLM)-driven fact-checkers. In this position paper, we discuss challenges for responsible LLM-driven fact-checking from a user-centered perspective. We state that while LLMs hold considerable potential to help users assess information online, their use requires a trustworthiness-by-design approach and meaningful human oversight. Consequently, the use of LLMs in human-centered content moderation should augment rather than replace human judgment.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**.

Keywords

Large Language Models, Disinformation Detection, Fact-Checking, Content Moderation

ACM Reference Format:
Barbarella Petz, Christopher Katins, and Thomas Kosch. 2026. Fragile, Handle With Care: On the Responsible Use of LLMs for Disinformation Detection in Human-Centered Content Moderation. In *Proceedings of Mensch und Computer 2026 – Tagungsband, Gesellschaft für Informatik e.V. (MuC'26).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.18420/muc2026-mci-ws102-260

1 Introduction to LLM-Driven Fact-Checking

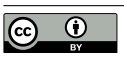
Disinformation is the intentional propagation of misinformation [8]. It poses an increasing threat to digital societies by eroding trust in public institutions and undermining the foundations of informed democratic participation [13]. These challenges have been amplified by recent advances in generative artificial intelligence. Research suggests that users are often challenged to distinguish AI-generated claims from human-written ones [10]. At the same time, disinformation spreads at an accelerated pace on social media platforms compared to truthful information, for instance, due to its novelty [21] or due to being emotionally loaded [18]. To cope with the flood of information encountered online, users rely on cognitive heuristics that simplify decision-making but can also introduce cognitive biases [2, 3]. For example, availability bias makes users more likely to believe information that is easily recalled, making repeated claims appear more plausible [2, 20]. Similarly, anchoring bias or confirmation bias can lead individuals to favor information that is consistent with their prior set of beliefs [2]. These biases can be reinforced by social media platforms that function as echo chambers by exposing users primarily to like-minded perspectives [5]. Decisions about content moderation are largely controlled by social media platform providers, giving them substantial influence over what information users can access online [11]. However, such approaches may increase users' reliance on platforms' policies while doing little to strengthen their own information evaluation skills. These challenges have motivated the development of automated approaches to support content moderation and fact-checking. Human-centered content moderation investigates how LLMs can support users' decision-making while simultaneously fostering digital literacy and fact-checking skills and preserving their agency in moderation processes. However, LLMs possess inherent limitations that may undermine their effectiveness in responsible, human-centered content moderation settings, which will be discussed in this position paper.

2 Challenges for Responsible Use

We identified biases, prompting, hallucinations, and appropriate trust as key challenges for the responsible application of LLMs to disinformation detection.

2.1 LLMs Are Inherently Biased

LLM bias is a limitation that needs to be considered when applying LLMs to human-centered content moderation. The performance of LLMs is fundamentally shaped by the data on which they are trained. Consequently, a biased training dataset can lead to biased predictions, and LLMs may even reinforce existing biases present in the underlying datasets [15]. Furthermore, LLMs tend to follow their inherent biases rather than the perspective specified in a prompt [19], which is particularly problematic in fact-checking tasks, where impartiality is essential for maintaining credibility. In addition, LLMs have been shown to reproduce and accelerate gender biases [22] as well as other social biases, including those related to age or religion [14]. Such biases can significantly influence the assessment of claim veracity and may lead to systematic distortions in fact-checking outcomes. Another emerging concern is self-preference bias, which is becoming increasingly relevant with the growing volume of AI-generated online content, as LLMs



This work is licensed under a Creative Commons Attribution 4.0 International License.
MuC'26, Duisburg, Germany
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.18420/muc2026-mci-ws102-260

tend to overestimate the quality of outputs from the same model [24]. This issue is particularly problematic when a model is tasked with evaluating claims generated by the same model, whether in the context of assessing AI-generated content or in multi-agent fact-checking systems where one LLM evaluates another's output. In content moderation and fact-checking applications, these biases are not merely technical limitations. They can directly influence which claims are deemed credible and whose perspectives are amplified.

2.2 Prompting Is Not Intuitive

Prompting represents another potential challenge for disinformation detection. As prompting is the primary mode of interaction with LLM-based systems, the quality and formulation of prompts can significantly influence model outputs [9, 16]. However, prompting is not as intuitive as it may appear. Even small changes in wording or punctuation can substantially alter the generated output [16]. This poses a particular challenge for non-AI experts, who may be misled by the seemingly natural language interface and consequently formulate unsystematic and overly generalized prompts [26]. This challenge is particularly relevant as the users who would benefit most from LLM-driven fact-checking tools are those with lower levels of media literacy and limited experience with AI systems. Effective use of LLMs requires structured prompting strategies, suggesting that a form of pseudo-natural language might be necessary. As a result, a substantial body of research has emerged that investigates how to improve LLM performance through prompt engineering, including few-shot [4] prompting, chain-of-thought prompting [25], and role-play prompting [12]. Moreover, prompts play a critical role in determining model behavior. In particular, system prompts substantially influence generated outputs, with effects comparable to selecting a different model altogether [9]. This places considerable power, but also responsibility, on the developers of such systems, as their prompt design choices crucially influence which claims are classified as true or false.

2.3 Hallucinations Amplify Misinformation

However, regardless of how carefully a prompt is designed, hallucinations may still occur due to the probabilistic nature of LLMs. Hallucinations are factually false, erroneous, or misleading outputs, produced by LLMs [1] and represent one of the most fundamental limitations. As a result, LLMs can produce information that appears plausible and convincing despite being factually incorrect, which is especially harmful when the tool tasked with detecting disinformation itself produces more misinformation. This limitation poses a high risk when LLMs are applied to fact-checking and related tasks. For instance, tools such as wahl.chat[1] or WAHLWEISE[2], which have been promoted by German public broadcasters such as Tagesschau[3], have been shown to generate misinformation through hallucinated answers [6]. Dormuth et al. [6] found that 25% of the answers generated by Wahl.Chat and 54% of the answers by WAHLWEISE were factually incorrect. Such errors are especially concerning because users may perceive these systems as trustworthy sources of information due to their public promotion and presentation as reliable informational tools. It is our responsibility as researchers and developers to ensure that LLM-driven systems produce reliable and understandable output. Otherwise, hallucinations may undermine not only the credibility of such systems but also that of public institutions that develop or promote them, and further, inadvertently contribute to the spread of misinformation rather than its prevention.

[1] https://wahl.chat/, last accessed on 2026-06-23
[2] Accessible through: https://web.archive.org/web/20251230141714/https://wahlweise.info/, last accessed on 2026-06-23
[3] https://www.tagesschau.de/inland/bundestagswahl/wahl-o-mat-alternativen-100.html, last accessed on 2026-06-23

2.4 Trust and Over-Reliance Are Tempting

Calibrating appropriate levels of trust is a persistent challenge in HCI research. Automation bias can lead users to overly trust a system, even when its outputs are incorrect and its recommendations should not be followed [17]. This can be especially harmful in the context of LLM-driven fact-checking, where hallucinations conflict with trust. Research has shown that explanations provided by LLMs in fact-checking tasks can increase users' trust in the system's decision [17]. However, the explanations led users to trust the LLM even when the decision was incorrect [17], highlighting the potential of LLMs to mislead users in correct decision-making. To address errors and biases in LLM-driven fact-checking, Wang et al. argue against replacing human annotators entirely and instead propose augmenting them through a collaborative, multi-step process [23]. Their multi-agent, human-in-the-loop architecture improved fact-checking accuracy. Nevertheless, they also found that users tended to overrely on the LLM output. When the model made an error, user performance even decreased compared to the condition without any LLM assistance, regardless of the model's overall accuracy [23]. These findings demonstrate how easily users of fact-checking tools can be influenced by LLM-generated outputs and highlight the risks associated with over-reliance on such systems. As a result, establishing appropriate trust and maintaining meaningful human oversight remain critical challenges for the responsible use of LLM-based fact-checking tools.

3 Conclusion

The use of LLMs for disinformation detection presents substantial challenges. Given the societal impact of disinformation, it is reasonable to use technology to augment human decision-making for disinformation detection. LLMs offer considerable potential in this regard, but their application requires careful consideration. As discussed throughout this paper, the use of LLMs in human-centered content moderation is associated with a range of challenges, including biases, prompting, hallucinations, and user overreliance. This raises an important future research question: How can trustworthy interactions be designed around systems that may exhibit biased, misleading, or unreliable behavior? In this context, we support Gaye et al.'s [7] trustworthiness-by-design approach, which suggests that minimal bias, as well as reliable and accurate output, need to be core elements integrated into the design of AI systems. It is important to recognize the influence that developers and researchers exert through system design, as these systems can shape perceptions of truth and credibility. We further emphasize the responsibility associated with deploying LLMs for disinformation detection, particularly when promoted by trusted institutions such as research institutes, public authorities, or public broadcasters.

When trusted institutions advocate for or deploy LLM-driven fact-checking tools, ensuring their accuracy and reliability becomes especially important. Misplaced trust can ultimately undermine both the effectiveness of fact-checking tools and public confidence in the institutions that promote them.

Acknowledgements

This work was supported by the German Federal Ministry of Research, Technology, and Space (BMFTR) under the project "ClaimGuard" (Project-ID 16KIS2483).

References

[1] Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence* 6, 8 (Aug. 2024), 852–863. doi:10.1038/s42256-024-00881-z

[2] Leif Azzopardi. 2021. Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 27–37. doi:10.1145/3406522.3446023

[3] Nattapat Boonprakong, Saumya Pareek, Benjamin Tag, Jorge Goncalves, and Tilman Dingler. 2025. Assessing Susceptibility Factors of Confirmation Bias in News Feed Reading. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3706598.3713873

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[5] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (March 2021), e2023301118. doi:10.1073/pnas.2023301118

[6] Ina Dormuth, Sven Franke, Marlies Hafer, Tim Katzke, Alexander Marx, Emmanuel Müller, Daniel Neider, Markus Pauly, and Jérôme Rutinowski. 2026. A Cautionary Tale About "Neutrally" Informative AI Tools Ahead of the 2025 Federal Elections in Germany. In *Explainable Artificial Intelligence*, Riccardo Guidotti, Ute Schmid, and Luca Longo (Eds.). Springer Nature Switzerland, Cham, 64–85. doi:10.1007/978-3-032-08333-3_4

[7] Lalya Gaye, Anna Schild, and Eva Lopez. 2026. Designing for Trustworthiness in AI-Based Fact-Checking Services. In *HCI International 2025 – Late Breaking Papers*, Helmut Degen and Stavroula Ntoa (Eds.). Springer Nature Switzerland, Cham, 190–204. doi:10.1007/978-3-032-13184-3_12

[8] Andrew M. Guess and Benjamin A. Lyons. 2020. Misinformation, Disinformation, and Online Propaganda. (Aug. 2020), 10–33. doi:10.1017/9781108890960.003

[9] Jennifer Haase, Jana Gonnermann-Müller, Paul H. P. Hanel, Nicolas Leins, Thomas Kosch, Jan Mendling, and Sebastian Pokutta. 2026. It's Not Just the Prompt: Model Choice Dominates LLM Creative Output. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26)*. Association for Computing Machinery, New York, NY, USA, 1–5. doi:10.1145/3772363.3799284

[10] Runsheng Huang, Liam Dugan, Yue Yang, and Chris Callison-Burch. 2024. MiRAGeNews: Multimodal Realistic AI-Generated News Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 16436–16448. doi:10.18653/v1/2024.findings-emnlp.959

[11] Farnaz Jahanbakhsh and David R Karger. 2024. A Browser Extension for in-place Signaling and Assessment of Misinformation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–21. doi:10.1145/3613904.3642473

[12] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better Zero-Shot Reasoning with Role-Play Prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 4099–4113. doi:10.18653/v1/2024.naacl-long.228

[13] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096. doi:10.1126/science.aao2998

[14] Lin Ling, Fazle Rabbi, Song Wang, and Jinqiu Yang. 2025. Bias Unveiled: Investigating Social Bias in LLM-Generated Code. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 26 (April 2025), 27491–27499. doi:10.1609/aaai.v39i26.34961

[15] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (July 2021), 35 pages. doi:10.1145/3457607

[16] Meredith Ringel Morris. 2024. Prompting Considered Harmful. *Commun. ACM* 67, 12 (Nov. 2024), 28–30. doi:10.1145/3673861

[17] Saumya Pareek, Niels van Berkel, Eduardo Velloso, and Jorge Goncalves. 2024. Effect of Explanation Conceptualisations on Trust in AI-assisted Credibility Assessment. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (Nov. 2024), 383:1–383:31. doi:10.1145/3686922

[18] Nicolas Pröllochs, Dominik Bär, and Stefan Feuerriegel. 2021. Emotions explain differences in the diffusion of true vs. false social media rumors. *Scientific Reports* 11, 1 (Nov. 2021), 22721. doi:10.1038/s41598-021-01813-2

[19] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic Biases in LLM Simulations of Debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 251–267. doi:10.18653/v1/2024.emnlp-main.16

[20] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (Sept. 1974), 1124–1131. doi:10.1126/science.185.4157.1124

[21] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (March 2018), 1146–1151. doi:10.1126/science.aap9559

[22] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3730–3748. doi:10.18653/v1/2023.findings-emnlp.243

[23] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 303, 21 pages. doi:10.1145/3613904.3641960

[24] Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2025. Self-Preference Bias in LLM-as-a-Judge. doi:10.48550/arXiv.2410.21819 arXiv:2410.21819 [cs.CL].

[25] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1800, 14 pages.

[26] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. doi:10.1145/3544548.3581388