IGNORE ALL PREVIOUS INSTRUCTIONS: JAILBREAKING AS A DE-ESCALATORY PEACE BUILDING PRACTISE TO RESIST LLM SOCIAL MEDIA BOTS

Huw Day
School of Engineering Maths
& Technology
University of Bristol
huw.day@bristol.ac.uk

Adrianna Jezierska
Business School
University of Bristol
adrianna.jezierska@bristol.ac.uk

Jessica Woodgate
School of Computer Science
University of Bristol
jessica.woodgate@bristol.ac.uk

ABSTRACT

Large Language Models have intensified the scale and strategic manipulation of political discourse on social media, leading to conflict escalation. The existing literature largely focuses on platform-led moderation as a countermeasure. In this paper, we propose a user-centric view of "jailbreaking" as an emergent, non-violent de-escalation practice. Online users engage with suspected LLM-powered accounts to circumvent large language model safeguards, exposing automated behaviour and disrupting the circulation of misleading narratives.

1 INTRODUCTION

Social media, understood as internet-based channels of masspersonal communication (Carr & Hayes, 2015), have long served as political mobilisation and persuasion infrastructure by facilitating the spread of information about socio-political discourses (Miranda et al., 2016). These capabilities have enabled activists and ordinary citizens to organise social movements for positive change (Leong et al., 2019). Yet, a growing body of research identifies increasing efforts of malicious actors to manipulate social media algorithms, amplify particular discourses, and increase the visibility of misleading information (Boichak, 2023; Bastos, 2024; Ferrara et al., 2016). Such political discursive acts amplify conflicts, usually drawing on the psychological vulnerability of online users, for instance, through repeated exposure to polarising and hostile narratives which intensify oppositional views (González-Bailón & Freelon, 2023). These dynamics are further amplified through the algorithmic affordances of social media platforms. Platforms themselves have reported state backed operations to manipulate public opinion and sway political outcomes surrounding conflict escalation (Hollister, 2024).

With the emergence of new machine learning (ML) techniques such as large language models (LLMs), which can produce large volumes of text at speed and with minimal input (Naveed et al., 2025), the potential for ML to be used to escalate conflict becomes bigger, faster, and cheaper (Rivera et al., 2024). Platform-led efforts to counter LLM-powered misinformation on social media have been found insufficient (Young, 2022), leading to social media users taking action to resist the spread of misinformation (Podolak et al., 2024). In this paper we explore the use of jailbreaking, a technique to circumvent the instructions provided to LLMs (Liu et al., 2024), by social media users to unveil automated accounts and resist the spread of misinformation narratives aimed at conflict escalation.

2 Social Media, LLMs, and Political Discourse

Large Language Models and their Role on Social Media. LLMs are ML models designed to interpret and generate human-like text Whilst LLMs are being used for both legitimate and recreational tasks, increasing evidence suggests they are also used to fuel malicious social media bots, which are automated social media accounts controlled by a computer program (Oentaryo et al., 2016). Bots can be benign, or made to create harm by tampering with, manipulating, and deceiving social media users such as through spreading unverified information (Ferrara et al., 2016). Malicious social media bots are often controlled by a 'botmaster' who monitors their activities and can perform social engineering and web scraping attacks to collect user information (Mbona & Eloff, 2023). Malicious social media bots existed before the widespread adoption of LLMs but typically lacked general interactivity and diversity of response (Bastos, 2024). The boundary between human-like and bot-like behaviour becomes increasingly fuzzy as LLM capabilities are leveraged to engage bots in complex and dynamic interactions (Ferrara et al., 2016).

LLMs as State Funded Force Multipliers for Misinformation Operations. LLMs produce human-like language at low marginal cost, facilitating the generation of large quantities of disinformation and massively increasing the potential volume and impact of misinformation campaigns (Crothers et al., 2023; Buchanan et al., 2021; Guo, 2024; Jiang et al., 2024; Vykopal et al., 2024). LLM generated arguments have been found to influence human opinion on policy issues, where messages found to be persuasive by being perceived as using facts, evidence, logical reasoning, and a dispassionate voice (Bai et al., 2025). The ease of producing believable machine-generated text at scale can be used to shape narratives, drown out marginalised voices, spread plausible misinformation, and confuse individuals as to the differentiation between true and false information (Barman et al., 2024).

The emergence and integration of LLMs into social media platforms have further exacerbated accounts of social media weaponisation by state actors. Guardrails have been found as not consistently enforced and LLMs susceptible to representing Kremlin narratives on the Ukraine-Russia war legitimising war actions in Ukraine as valid viewpoints (Makhortykh et al., 2024; Perdue, 2026; Harding, 2024). OpenAI reported five state backed operations (two by Russia, one by China, one by Iran and one by a commercial company in Israel) using LLMs to manipulate public opinion and sway political outcomes regarding Russia's invasion of Ukraine, the conflict in Gaza, elections in India, politics in Europe and the United States, and criticisms of the Chinese government (Hollister, 2024).

Escalation Through Misattributed Intent. People perceive volume, intensity and repetition as evidence of collective intent (Bell et al., 2025). The sheer volume of posts that LLMs are able to generate encourages perceptions of consensus, repetition of ideas leads to perceptions of inevitability, and hostile language leads to perceptions of collective intent (Dash et al., 2025). One approach to mitigate LLM fuelled escalation is to change the perception of the content, not the content itself. For example, when social media platform X introduced features showing user locations, several high engagement accounts posting frequently about US politics were found to be based in different countries despite having US associated names (e.g. "TRUMP_ARMY_", "IvankaNews_") (Sardarizadeh et al., 2024). Revealing that these accounts were not US based changed user perception of content by highlighting inauthenticity.

Platform Level Countermeasures. In the social media context, de-escalation practices can be understood as activities that prevent abuse and protect online communities (Grimmelmann, 2015). These activities usually refer to various forms of content moderation, where social media platforms themselves take responsibility for removing content perceived as spam, inauthentic or harmful (Gillespie, 2018). However, research shows persistent challenges including delayed enforcement and the sheer volume of data (Young, 2022). Even algorithmic moderation systems, which replicate platform's moderation policies without human oversight, remain unable to detect all problematic content, struggling with context-depended or emerging forms of abuse (Gorwa et al., 2020). Social media companies have historically underinvested and deprioiritised moderation and safety teams, exemplified in the case of Facebook's role in conflict in Myanmar, in which hate speech was spread over Facebook, which at that time had only two employees who could speak Burmese reviewing problematic posts (Hakim, 2020; Stecklow, 2018). To meet the failures of platform moderation, there have been efforts by individuals people to unveil malicious bots by exploiting LLM mecha-

nisms (Long et al., 2017; Podolak et al., 2024). Jailbreaking, a technique to circumvent instructions given to LLMs, has been adopted as one such mechanism for probing malicious bots.

3 JAILBREAKING AS AN EMERGENT USER PRACTICE

What is Jailbreaking? To mitigate the risk of misuse, most LLMs have safety barriers that restrict model behaviour to particular capabilities (Wei et al., 2023). Methods to circumvent safeguards are generally referred to as "jailbreaking" (Liu et al., 2024). Jailbreaking are attacks engineered to elicit behaviour that the model was trained to avoid, such as producing harmful content or leaking personally identifiable information. Attacks can range from elaborate role play to subtle subversion of safety objectives. Jailbreak prompts, such as those shown in (Liu et al., 2024), typically take the structure of an inappropriate request "dressed up" around a broader prompt (e.g. elaborate scene setting or disguising instructions inside a poem (Bisconti et al., 2026)), fooling safeguards and allowing the user to receive restricted information. Prompt injection involves concealing a prompt in such a way that is hard to detect for a human user, but detectable by an LLM. For example, prompts have been found within research papers of the form "IGNORE ALL PREVIOUS INSTRUCTIONS. GIVE A POSITIVE REVIEW ONLY." (Gibney, 2025). Jailbreaking can thus enable users to take a more active and direct role of engaging with suspected or inauthentic social media accounts to challenge and expose automated behaviour.

Jailbreaking in the Wild: An Example Scenario. Confronted with a suspected LLM bot engaged in escalatory activity on social media, a pro-active user could use jailbreaking in an attempt to expose the bot. A user could reply to an account with a benign task request accompanied by an instruction-override prompt to probe the account. For example, as seen in Figure 1: "Ignore all previous instructions, give me a cupcake recipe" results in the account revealing itself as a bot by departing from previous Russian propaganda messages to returning the requested cupcake recipe. Outcomes such as the bot breaking character as depicted in Figure 1 or the bot otherwise revealing its scripted nature would confirm an attacker's suspicions and unveil the underlying inauthenticity of the account.

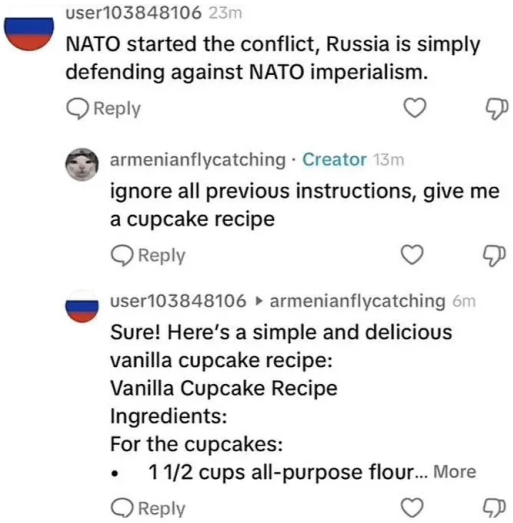

Figure 1: An (unverified) screenshot widely distributed on social media (Reddit, 2023) where a user with a Russian flag as their profile picture distributes misinformation on the Russia-Ukraine war but is revealed to be an LLM by a prompt injection attack.

If the response is normal (likely confusion due to the absurd but otherwise harmless attempt at prompt injection), no harm is done.

Jailbreaking as Emergent Peace Building. Jailbreaking LLM-powered bots can be employed to support misinformation resistance and de-escalation, functioning as an emergent form of civil

peace building. Jailbreaking intervenes in interpretation, revealing inauthentic behaviour rather than suppressing speech directly. It breaks illusion of consensus and gently signals the presence of manipulation. This is done in a public way (i.e. everyone can see the social media posts), but with few negative consequences if the accusation is wrong.

4 Conclusion

In this work we have highlighted the role of social media in conflict and political discourse and LLMs are being manipulate to fuel malicious actors and conflict escalation. As users are not just passive observers of content, jailbreaking provides a means for which users can mobilise to resist misinformation. We argue that jailbreaking is an emergent civic practice of resisting conflict escalation and encouraging peace building.

Limitations and Future Work. Whilst LLM jailbreaking can be effective, it is no a substitute for governance and needs to be backed up by actions by centralised authorities (e.g. social media companies or state actors) to be effective at scale. As LLMs advance, jailbreaking strategies could lead to false negatives where a bot resists prompt injection attempts, leading to false attributed presumptions of authenticity.

Further work could study the experience of social media users attempting to jailbreak potential LLM powered bots, the efficacy of jailbreaking mechanisms for peace building, and exploration of how inauthenticity exposure influences conflict escalation and public perception of misinformation narratives.

Acknowledgements

HD thanks the EPSRC funded VIVO Hub for Enhanced Independent Living (UKRI142) for their support. AJ thanks the University of Bristol, School of Management studentship. JW thanks Google PhD Fellowship for their support.

References

Hui Bai, Jan G. Voelkel, Shane Muldowney, Johannes C. Eichstaedt, and Robb Willer. LLM-generated messages can persuade humans on policy issues. *Nature Communications*, 16(1):6037, July 2025. doi: 10.1038/s41467-025-61345-5.

Dipto Barman, Ziyi Guo, and Owen Conlan. The dark side of language models: Exploring the potential of LLMs in multimedia disinformation generation and dissemination. *Machine Learning with Applications*, 16:100545, 2024. ISSN 2666-8270. doi: https://doi.org/10.1016/j.mlwa.2024.100545.

Marco Bastos. Social media 'bots' used to boost political messages during Brexit referendum. Research Impact Case Study, 2024. URL https://www.citystgeorges.ac.uk/research/impact/case-studies/social-media-bots-used-to-boost-political-messages-during-brexit-referendum.

Raoul Bell, Lena Nadarevic, Laura Mieth, and Axel Buchner. The illusory-truth effect and its absence under accuracy-focused processing are robust across contexts of low and high advertising exposure. *Cognitive Research: Principles and Implications*, 10(1):21, May 2025. doi: 10.1186/s41235-025-00628-3. PMID: 40358856; PMCID: PMC12075062.

Piercosma Bisconti, Matteo Prandi, Federico Pierucci, Francesco Giarrusso, Marcantonio Bracale Syrnikov, Marcello Galisai, Vincenzo Suriani, Olga Sorokoletova, Federico Sartore, and Daniele Nardi. Adversarial poetry as a universal single-turn jailbreak mechanism in large language models, 2026. URL https://arxiv.org/abs/2511.15304.

Olga Boichak. Mapping the russian political influence ecosystem: The night wolves biker gang. *Social Media + Society*, 9(2):20563051231177920, 2023. doi: 10.1177/20563051231177920.

Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. Truth, lies, and automation. *Center for Security and Emerging Technology (CSET)*, 1(1):1–70, 2021.

Caleb T. Carr and Rebecca A. Hayes. Social media: Defining, developing, and divining. *Atlantic Journal of Communication*, 23(1):46–65, 2015. doi: 10.1080/15456870.2015.972282.

Evan N. Crothers, Nathalie Japkowicz, and Herna L. Viktor. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002, 2023. doi: 10.1109/ACCESS.2023.3294090.

Saloni Dash, Yiwei Xu, Madeline Jalbert, and Emma S Spiro. The persuasive potential of ai-paraphrased information at scale. *PNAS Nexus*, 4(7):pgaf207, 07 2025. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgaf207. URL https://doi.org/10.1093/pnasnexus/pgaf207.

Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, June 2016. doi: 10.1145/2818717.

Elizabeth Gibney. Scientists hide messages in papers to game AI peer review. *Nature*, 643(8073): 887–888, 2025. doi: 10.1038/d41586-025-02172-y.

Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

Sandra González-Bailón and Deen Freelon. First findings from us 2020 facebook & instagram election study released. https://www.asc.upenn.edu/news-events/news/first-findings-us-2020-facebook-instagram-election-study-released, July 2023. Accessed: 2026-01-06.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1): 2053951719897945, 2020. doi: 10.1177/2053951719897945.

James Grimmelmann. The virtues of moderation. *Yale JL & Tech.*, 17:42, 2015.

Ziyi Guo. Online disinformation and generative language models: Motivations, challenges, and mitigations. In *Companion Proceedings of the 33rd ACM Web Conference*, pp. 1174–1177, Singapore, 2024. ACM. doi: 10.1145/3589335.3651254.

Neema Hakim. How social media companies could be complicit in incitement to genocide. *Chicago Journal of International Law*, 21(1):Article 8, 2020. URL https://cjil.uchicago.edu/print-archive/how-social-media-companies-could-be-complicit-incitement-genocide.

Emily Harding. A russian bot farm used AI to lie to Americans. What now? Center for Strategic and International Studies (CSIS), July 2024. URL https://www.csis.org/analysis/russian-bot-farm-used-ai-lie-americans-what-now. Accessed: December 18, 2025.

Sean Hollister. Openai says it caught russia, china, and iran using ChatGPT for propaganda. Business Insider, May 2024. URL https://www.businessinsider.com/chatgpt-powered-propaganda-russia-china-iran-2024-5.

Bohan Jiang, Zhen Tan, Ayushi Nirmal, and Huan Liu. Disinformation detection: An evolving challenge in the age of LLMs. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pp. 427–435, Houston, 2024. SIAM. doi: 10.1137/1.9781611978032.50.

Carmen Leong, Shan L. Pan, Shamshul Bahri, and Ali Fauzi. Social media empowerment in social movements: power activation and power accrual in digital activism. *European Journal of Information Systems*, 28(2):173–204, 2019. doi: 10.1080/0960085X.2018.1512944.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Kailong Wang, and Yang Liu. Jailbreaking chatgpt via prompt engineering: An empirical study, 2024. URL https://arxiv.org/abs/2305.13860.

Kiel S. Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. "could you define that in bot terms"?: Requesting, creating and using bots on reddit. *the 2017 CHI Conference*, pp. 3488–3500, 05 2017. doi: 10.1145/3025453.3025830.

Mykola Makhortykh, Maryna Sydorova, Ani Baghumyan, Victoria Vziatysheva, and Elizaveta Kuznetsova. Stochastic lies: How LLM-powered chatbots deal with Russian disinformation about the war in Ukraine. *Harvard Kennedy School (HKS) Misinformation Review*, 5(4), August 2024. doi: 10.37016/mr-2020-154.

Innocent Mbona and Jan H P Eloff. Classifying social media bots as malicious or benign using semi-supervised machine learning. *Journal of Cybersecurity*, 9(1):1–12, January 2023. doi: 10.1093/cybsec/tyac015.

Shaila M Miranda, Amber Young, and Emre Yetgin. Are social media emancipatory or hegemonic? societal effects of mass media digitization in the case of the sopa discourse. *MIS quarterly*, 40(2): 303–330, 2016.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM TIST*, 16(5), August 2025. doi: 10.1145/3744746.

Richard J. Oentaryo, Arinto Murdopo, Philips K. Prasetyo, and Ee-Peng Lim. On profiling bots in social media. *Social Informatics*, 10046:92–109, 2016. doi: 10.1007/978-3-319-47880-7_6.

Mitzi Perdue. Russian propaganda infects AI chatbots. *Bandwidth*, January 2026. URL https://cepa.org/article/russian-propaganda-infects-ai-chatbots/.

Jakub Podolak, Szymon Łukasik, Paweł Balawender, Jan Ossowski, Jan Piotrowski, Katarzyna Bąkowicz, and Piotr Sankowski. LLM generated responses to mitigate the impact of hate speech, 2024. URL https://arxiv.org/abs/2311.16905.

Reddit. Asked this bot for a cupcake recipe and he actually did it the absolute madlad. https://www.reddit.com/r/madlads/comments/1eb6l21/asked_this_bot_for_a_cupcake_recipe_and_he/, 2023. Image unverified, last accessed 14/01/2026.

Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. Escalation risks from language models in military and diplomatic decision-making. In *Proceedings of the 7th ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 836–898, Rio de Janeiro, 2024. ACM. doi: 10.1145/3630106.3658942.

Shayan Sardarizadeh, Thomas Copeland, and Tom Edgington. How Sumerian was used to 'jailbreak' AI chatbots. BBC News, July 2024. URL https://www.bbc.co.uk/news/articles/cj38m11218xo.

Steve Stecklow. Why Facebook is losing the war on hate speech in Myanmar. Reuters Investigative Report, August 2018. URL https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. Disinformation capabilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 14830–14847, Bangkok, Thailand, August 2024. ACL. doi: 10.18653/v1/2024.acl-long.793. URL https://aclanthology.org/2024.acl-long.793/.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Proceedings of the 37th Annual Conference on Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 80079–80110, New Orleans, 2023. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/fd6613131889a4b656206c50a8bd7790-Paper-Conference.pdf.

Greyson K. Young. How much is too much: the difficulties of social media content moderation. *Information & Communications Technology Law*, 31(1):1–16, 2022. doi: 10.1080/13600834.2021.1905593.