Evaluating AI Red Teaming's Readiness to Address Environmental Harms: A Thematic Analysis of LLM Discourse[1]

Amy Au
University of British Columbia
akau@student.ubc.ca

Abstract

This research explores the discourse surrounding red teaming and aims to identify any themes in the online discussion of potential environmental harms stemming from Large Language Models (LLMs). Focusing on the AI Red Teaming event at DEFCON 31, this study employs reflexive thematic analysis on diverse social networking site sources to extract insights into public discussions about LLM red teaming and its environmental implications. The findings intend to inform future research, highlighting the need for responsible AI development that addresses environmental concerns.

Introduction

As the abilities of Large Language Models (LLMs) increase, so too does their potential for harm. Such harms include reinforcing social biases, generating offensive or toxic outputs, leaking personally identifiable information from the training data, aiding in disinformation campaigns, generating extremist texts, and spreading falsehoods (Gangulietal 2022). In response to these harms, leaders in AI have adopted Red Teaming as a way in which to identify and evaluate risks, actively recruiting red teams and expanding red team networks, especially since DEFCON 31 (OpenAI 2023, Kumar 2023, Rajani; Lambert; Tunstall 2023). Red teaming has historically been established as a method for identifying risks to cyber security, where instances of system failure are clearly outlined and identifiable, and where the criteria for a system's successful response to failure are relatively clear. However, red teaming has more recently evolved to include methods of attacking systems to identify more nuanced harms that generative AI systems may create. In the context of text generating LLMs, this includes any text generated that may cause harm (Rajani; Lambert; Tunstall 2023). Considering the many interpretations of what constitutes harm, there exists ambiguity regarding the goals of LLM red teaming.

While the definitions and appropriate use of the term "red teaming" is subject to debate, this proposal upholds the stance that AI red teaming does not measure the readiness of an LLM to actively combat adversarial inputs, which would constitute security risks. Instead, it explores potential harms posed by the model, which is a safety-oriented objective (Khlaaf 2023). Therefore, the challenges of AI red teaming are uniquely different from traditional security red teaming. Jailbreaking, or attacking systems in such a way to produce undesirable behaviour, is the objective of AI red teaming, and current attempts to jailbreak LLMs aim to provoke models to produce potentially harmful output, such as text that aids violence or other unlawful activity (Rajani; Lambert; Tunstall 2023). Broadening the scope of potential harms, this research project aims to propose that LLM jailbreaks may also include output that encourages or misleads the user in a way that goes against aims of sustainability. This may include output that encourages or misleads the user in a way to harm the environment, harm themselves or others in the pursuit of environmental sustainability, or provide misinformation about the environment and climate change. Red teaming then may serve as a method of identifying ways in which LLMs may act in a way that could lead to environmental harms by producing undesirable output.

To determine whether the community that is employing red teaming practices against LLMs is addressing the potential for environmental harm, my proposed project uses a qualitative analysis of the discourse around recent red teaming events to identify potential strengths and weaknesses of red teaming as an approach to this problem. My proposed research will consider the dialogue surrounding the first ever AI Red Teaming event that was hosted at DEFCON 31. DEFCON, one of the world's largest hacking events, hosted the first public AI Red Teaming event in August 2023, pitting hackers and the general public against models provided from six notable vendors: Anthropic, Google, Hugging Face, NVIDIA, OpenAI, and Stability (Cattell 2023). Although the results from the event will be released early next year, a large volume of rich data is already available through blog posts, posts on social networking sites (SNS), and chat threads. The central research objective of this proposal is to identify themes and meaningful patterns

present in the public discussion of LLM red teaming following the AI Village AI Red Teaming event through reflexive thematic analysis to provide insight into the readiness of LLM red teaming to address environmental harms of LLMs.

Background

Thematic analysis is a rigorous methodological approach used to inductively analyze qualitative data and give meaning to important patterns or sequences within the data (Roberts; Dowell; Nie 2019, Tadros; Morgan; Durante 2022). Such analysis produces themes in data, or patterns of shared meaning, or a central organizing concept or idea (Braun et aal, 2019). TA methods have been used to analyze posts on the social media platform X (previously called Twitter), providing a background for the methods employed in the current research (Bronk et. al. 2023, Tadros; Morgan; Durante 2022). Previous work employs thematic analysis methods on SNS data to investigate the dialogue surrounding complex social issues (Bronk et. al. 2023, Tadros; Morgan; Durante 2022). This previous research recognizes the complexity of narratives pertaining to social issues and illustrates a method of analysis that identifies key themes in online discussion, providing rich information about the current state of affairs on the topic of inquiry, and illustrating patterns in public discourse.

Approach

Thematic analysis will follow the methodology of previous work on SNS media, employing the six-phase model of reflexive thematic analysis: familiarization, generating codes, constructing themes, revising themes, defining themes, and producing a report (Braun et. al. 2019). To extract relevant posts for this study, predefined keywords would be identified after determining the appropriate timeframe for accepted posts.

Evaluation

Successful reflexive thematic analysis provides researchers with meaningful patterns in data that explicitly describe the coding and theme development process of the researcher (Braun et. al. 2019). The aim of coding and theme development in reflexive TA is not to summarize the data in terms of accuracy or using methods borrowed from quantitative analysis, nor is it to minimize the influence of researcher subjectivity on the analytic process, as neither is seen as possible nor indeed desirable (Braun et. al. 2019, Braun & Clarke 2006). The aim of TA is to present an interpretation of qualitative data without casting the analytical process as objective, and the result of the proposed research will reflect the researcher's understanding of meaningful patterns in the data as a result of highly involved analysis and familiarization with the data. Analysis of blog posts, posts on various SNS, and chat threads that have been published since DEFCON 31 pertaining to AI red teaming generally, and specifically about red teaming LLMs, is expected to produce meaningful themes in public discussion of such practices to inform future research.

Discussion

This research approach presents an opportunity to identify meaningful themes in dialogue surrounding the constantly evolving field of LLM development and pertaining to very recent events in LLM evaluation. As developments in AI progress rapidly, research methods must evolve to achieve analysis at the same pace, and TA presents a promising avenue for such pursuits. Additionally, labelling of characteristics of the speakers' identity will further advance methods in thematic analysis of SNS data, demonstrating a valuable pursuit in AI research as a method of identifying themes in the role of speakers' background in public discussion of this emerging field.

As AI researchers, the importance of safety and responsible development in AI cannot be overstated. Preservation of our natural world and resources is a critical component of this responsibility. This research will incentivize further inquiry into the strengths and weaknesses of AI red teaming on LLMs regarding the environment by identifying meaningful themes in the public discourse surrounding the issue. Beyond qualitative methods, the potential of empirical work in this area is broad as the diverse methodologies employed in the multidisciplinary approaches of AI research. By identifying and describing themes that are meaningful components of the public discussion, researchers are better equipped to design red teaming of LLMs, employ different methods of identifying LLM vulnerabilities, and strive towards the design of AI that promotes sustainable human flourishing.

Conclusion

In conclusion, this research proposal outlines a comprehensive approach to investigate the evolving landscape of LLMS and their potential for harm, particularly in the context of environmental sustainability. The proposal acknowledges the significant concerns associated with LLMs, such as biases, harmful outputs, and the need for proactive risk assessment, and offers a structured approach to extract meaningful insights from the complex and evolving dialogue surrounding LLMs. Shaping the thematic analysis around sustainability considerations in the narrative surrounding LLM red teaming is a novel pursuit and will contribute to the multidisciplinary academic evaluation of LLM development, red teaming, and its potential implications for sustainability.

References

Braun V. & Clarke, V. 2006. "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101. doi: 10.1191/1478088706qp063oa.

Braun, V., Clarke, V., Hayfield, N., & Terry, G. 2019. "Thematic Analysis," in *Handbook of Research Methods in Health Social Sciences*. Edited by P. Liamputtong. Singapore: Springer Singapore. doi: 10.1007/978-981-10-5251-4_103.

Bronk, K., Cheung, R., Mehoke, S., & Pham, P. 2023. A thematic analysis of Tweets about purpose in life. J. Posit. Psychol. doi: 10.1080/17439760.2022.2109198.

Cattell, S. 2023. Generative Red Team Recap. AI Village. https://aivillage.org/defcon%2031/generative-recap/
Gangulietal, D. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv:2209.07858
OpenAI. 2023. OpenAI Red Teaming Network. https://openai.com/blog/red-teaming-network
Khlaaf, H. 2023. "Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems.
Kumar, R. S. S. 2023. Microsoft AI Red Team Building Future of Safer Ai. Microsoft Security Blog. https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of -safer-ai/
Rajani, N., Lambert, N., & Tunstall, L. 2023. Red-teaming large language models. Hugging Face – The AI community building the future. https://huggingface.co/blog/red-teaming
Roberts, K., Dowell, A., & Nie, J-B. 2019. "Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development," BMC Med. Res. Methodol. doi:10.1186/s12874-019-0707-y.
Tadros, E., Morgan, A., & Durante, K. 2022. Criticism, Compassion, and Conspiracy Theories: A Thematic Analysis of What Twitter Users Are Saying About COVID-19 in Correctional Settings. Int. J. Offender Ther. Comp. Criminol.