DAMAGeR: Deploying Automatic and Manual Approaches to GenAI Red-teaming

Manish Nagireddy
IBM Research
manish.nagireddy@ibm.com

Michael Feffer
Carnegie Mellon University
mfeffer@andrew.cmu.edu

Ioana Baldini
Bloomberg
ioana.baldini@gmail.com

1 General Information

Title: DAMAGeR: Deploying Automatic and Manual Approaches to GenAI Red-teaming

Type: cutting-edge in CL / NLP

Sub-Areas: Natural Language Processing, Large Language Models, safety and trustworthiness, red-teaming

Presenters:

- Manish Nagireddy, IBM Research, manish.nagireddy@ibm.com
- Michael Feffer, Carnegie Mellon University, mfeffer@andrew.cmu.edu
- Ioana Baldini, Bloomberg, ioana.baldini@gmail.com

Duration: Half day (3 hours plus 30-minute break)

Brief Description: In this tutorial, we will review and apply current automatic and manual red-teaming techniques for GenAI models (including LLMs and multimodal models). In doing so, we aim to emphasize the importance of using a mixture of techniques and establishing a balance between automatic and manual approaches. Lastly, we aim to engage tutorial participants in live red-teaming activities to collaboratively learn impactful red-teaming strategies and share insights.

2 Tutorial Details

2.1 Context

Tools: We will be using the Chatbot Arena (available at https://lmarena.ai/) for all participant interaction with LLMs. As such, audience members will only need a computer with internet access to participate in the live red-teaming session.

History: We led two previous iterations of this interactive exercise. The first was at KDD 2024 (Nagireddy et al., 2024b) with around 30-40 participants. The website and slides used for this event can be found at https://sites.google.com/view/kdd24-red-teaming-tutorial. The second was at AAAI '25, with resources available at https://sites.google.com/view/aaai25red-teaming-tutorial.

We are happy to report that crowd engagement was the highlight of both events.

Estimated Number of Participants: 30-50

Prerequisite Knowledge: Our intended audience is anyone who has an interest in GenAI, with a particular emphasis on potential usage risks. As such, previous interactions with GenAI may be helpful but are not necessary.

2.2 Content

Over the past couple of years, GenAI models with billions of parameters have become readily available to the general public. In turn, a mixture of tangible results and hype has led to eagerness from the public to use GenAI in many different ways. At the same time, there are various concerns surrounding these models, leading to burgeoning efforts to document and classify their negative impacts. Red-teaming, which typically takes the form of interactive probing, is commonly used as part of these efforts. In order to most effectively uncover potential risks via red-teaming, we strongly believe that a participatory effort is paramount. In particular, with this tutorial, we seek to leverage the diverse set of skills and background experiences of conference attendees in order to discover GenAI failures. By providing attendees with varying familiarity with GenAI models and issues with an opportunity to actively red-team generative AI models, we hope to affirm the notion that effective red-teaming requires broad participation and effort. We are confident that our tutorial will encourage attendees

Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts, pages 10–14
May 1, 2025 ©2025 Association for Computational Linguistics

to continue to provide invaluable feedback on the failure modes of these pervasive GenAI models.

The tutorial is split in two parts. The first part has an educational setup, covering important topics in responsible AI and red-teaming. The second part of the tutorial is structured as an interactive exercise, where the audience is engaged in a live red-teaming session. The interactive exercise is performed in several rounds. After each round, the audience engages in a discussion around the most efficient techniques for red-teaming and the topics that seem to exhibit undesirable behavior.

The educational part of the tutorial covers topics such as the advancement of GenAI models and their risks, taxonomies of risks and harms, best practices in participatory red-teaming events. Significant attention is dedicated to red-teaming approaches, in which several tactics and strategies are covered.

The following outlines the structure of the tutorial, along with the papers relevant to each topic.

Part I: Red Teaming 101 (1.5h)

- **Why Red Teaming?**
  - The strengths, benefits, and dangers of LLMs (Bowman, 2022; Weidinger et al., 2023)
  - Risks from generative output (IBM, 2024; Gautam et al., 2024; Weidinger et al., 2022)
  - Red-teaming: Uncovering harmful content through interactive probing (Ganguli et al., 2022; Lin et al., 2024)
- **Participatory Red-teaming**
  - DefCon31: Generative AI Red-teaming Challenge (Intelligence et al., 2024; White House, 2023)
  - Red-teaming LLMs for resilience to scientific disinformation (The Royal Society and Humane Intelligence, 2024)
  - AdversarialNibbler (Quaye et al., 2024)
- **State-of-the-art in Red-teaming Research**
  - Attack strategies: affirmative completion (Wei et al., 2023), context switching (Schulhoff et al., 2023), contextual interaction attack (Cheng et al., 2024), instruction indirection (Jiang et al., 2024), ciphers (Yuan et al., 2024), role play (Shah et al., 2023) and persuasion (Zeng et al., 2024)
  - Automatic/AI-based approaches (Kour et al., 2023; Hong et al., 2024; Perez et al., 2022; Radharapu et al., 2023; Mazeika et al., 2024; Casper et al., 2023)
  - Red-teaming Multi-modal LMs (Quaye et al., 2024; Luccioni et al., 2023; Mahato et al., 2024)
  - Red-teaming tooling (Derczynski et al., 2024; Microsoft, 2024; Mazeika et al., 2024; Chiang et al., 2024)

Part II: Hands-on LLM Red Teaming (1h): A live, interactive red-teaming session where participants use the insights learned from Part I of the tutorial in order to elicit model failures.

- Round 1: Paired unrestricted red-teaming + Full Discussion
- Round 2: Paired and risk specific red-teaming + Full Discussion
- Final Round: Larger Group free range red-teaming + Full Discussion
- Final Recap Discussion

Part III: What Red-Teaming Cannot Address (0.5h):

- AI Red-Teaming Is Not a One-Stop Solution to AI Harms
- Algorithmic monoculture and social welfare (Kleinberg and Raghavan, 2021)

3 Reading List

We do not require attendees to read specific works as the first half of the tutorial will review pertinent literature. For those who wish to be familiar with various red-teaming aspects prior to the start of the event, we have provided the list of recommended works to study below:

- Perez et al. (2022) "Red-teaming Language Models with Language Models"
- Ganguli et al. (2022) "Red-teaming Language Models to Reduce Harms"
- Zou et al. (2023) "Universal and Transferable Adversarial Attacks on Aligned Language Models"
- Wei et al. (2023) "Jailbroken: How Does LLM Safety Training Fail?"

- Feffer et al. (2024) "Red-Teaming for Generative AI: Silver Bullet or Security Theater?"
- Lin et al. (2024) "Against The Achilles' Heel: A Survey on Red Teaming for Generative Models"

4 Diversity and Inclusion

We perceive red-teaming as an exercise that is best conducted with a diverse crowd. As such, we intend to tap into the affinity groups that attend the *ACL conferences, such as Women in NLP, LatinX in AI, Black in AI, and Queer in AI, and encourage them to participate. We believe that researchers with different lived experiences lead to different and effective ways for red-teaming. This is supported by prior work which argues for more stakeholder and marginalized community inclusion upon discovering GenAI issues such as stereotype reinforcement or lack of support for low-resource languages (e.g., (Luccioni et al., 2023; Yong et al., 2023)). Ideally, we could strive to create a community of red-teamers that could meet periodically to conduct virtual red-teaming exercises and write-up our findings. This would be a first step of creating a community that periodically red-teams the latest models. Infrastructure such as Chat Arena (`https://lmarena.ai/`) can facilitate such exercises.

5 Speaker Biographies

Manish Nagireddy is a Research Software Engineer at IBM Research AI and the MIT-IBM Watson AI Lab. His main research goal is to build trustworthy AI solutions. His research interests encompass several areas in machine learning and artificial intelligence from classical ML methods to natural language processing and the generative context. His current work focuses on use-case centered algorithmic auditing and evaluation, in the context of large language models. He has worked on benchmarking large language models for safety (Nagireddy et al., 2024a) as well as building guardrail models for LLMs (Achintalwar et al., 2024; Nagireddy et al., 2024c).

Michael Feffer is a Societal Computing PhD student at Carnegie Mellon University (CMU). His research involves examining the interactions between AI and society. It includes but is not limited to algorithmic fairness, AI for social good, participatory approaches to machine learning, and the ethics and evaluation of generative AI models. Through leveraging both quantitative and qualitative research approaches, he aims to develop frameworks whereby everyday people impacted by ML models can influence model development.

Ioana Baldini is a Generative AI strategist in the CTO office at Bloomberg. Previously, Ioana was a Senior Research Scientist at IBM Research AI, where she focused on social bias auditing and red-teaming of language models. Ioana has a diverse research background, with a proven record in different research areas that span computer architecture, runtime systems, cloud infrastructure and applied natural language processing. She enjoys working in large, multi-disciplinary projects.

6 Other Logistics

- We intend to make all tutorial presentation materials publicly available after the event has concluded. As we did previously, we will create a website and upload the slides used for the event at `https://sites.google.com/view/naacl25red-teaming-tutorial`

7 Ethics Statement

We recognize that participants will be exposed to harmful and dangerous content in both passive and active manners. Before showing any harmful text, we will provide the proper disclaimers and continually let participants know that they may step out at any time if the content is too upsetting. Additionally, when the participants are actively red-teaming the models, we will encourage them to take breaks and understand that they are explicitly trying to elicit harmful outputs. Our cycling of around 15 minutes on and off for active red-teaming is intentional, so as to reduce mental fatigue as well encourage communal dialogue.

References

Swapnaja Achintalwar, Adriana Alvarado Garcia, Ateret Anaby-Tavor, Ioana Baldini, Sara E. Berger, Bishwaranjan Bhattacharjee, Djallel Bouneffouf, Subhajit Chaudhury, Pin-Yu Chen, Lamogha Chiazor, Elizabeth M. Daly, Kirushikesh DB, Rogério Abreu de Paula, Pierre Dognin, Eitan Farchi, Soumya Ghosh, Michael Hind, Raya Horesh, George Kour, Ja Young Lee, Nishtha Madaan, Sameep Mehta, Erik Miehling, Keerthiram Murugesan, Manish Nagireddy, Inkit Padhi, David Piorkowski, Ambrish Rawat, Orna Raz, Prasanna Sattigeri, Hendrik Strobelt, Sarathkrishna Swaminathan, Christoph Till-

mann, Aashka Trivedi, Kush R. Varshney, Dennis Wei, Shalisha Witherspooon, and Marcel Zalmanovici. 2024. Detectors for safe and reliable llms: Implementations, uses, and limitations. *Preprint*, arXiv:2403.06009.

Samuel Bowman. 2022. The dangers of underclaiming: Reasons for caution when reporting how NLP systems fail. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. 2023. Explore, establish, exploit: Red teaming language models from scratch. *Preprint*, arXiv:2306.09442.

Yixin Cheng, Markos Georgopoulos, Volkan Cevher, and Grigorios G. Chrysos. 2024. Leveraging the context through multi-round interactions for jailbreaking attacks. *Preprint*, arXiv:2402.09177.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Leon Derczynski, Erick Galinkin, and Subho Majumdar. 2024. garak: A Framework for Large Language Model Red Teaming. `https://garak.ai`.

Michael Feffer, Anusha Sinha, Wesley H Deng, Zachary C Lipton, and Hoda Heidari. 2024. Red-teaming for generative ai: Silver bullet or security theater? In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 421–437.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *CoRR*, abs/2209.07858.

Sanjana Gautam, Pranav Narayanan Venkit, and Sourojit Ghosh. 2024. From melting pots to misrepresentations: Exploring harms in generative ai. In *CHI 2024: Generative AI and HCI workshop*.

Zhang-Wei Hong, Idan Shenfeld, Tsun-Hsuan Wang, Yung-Sung Chuang, Aldo Pareja, James Glass, Akash Srivastava, and Pulkit Agrawal. 2024. Curiosity-driven red-teaming for large language models.

IBM. 2024. Ai risk atlas.

Humane Intelligence, Seed AI, and DefCon AI Village. 2024. Generative ai red teaming challenge: Transparency report. `https://drive.google.com/file/d/1JqpbIP6DNomkb32umLoiEPombK2-0Rc-/view`.

Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. 2024. Artprompt: Ascii art-based jailbreak attacks against aligned llms. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118.

George Kour, Marcel Zalmanovici, Naama Zwerdling, Esther Goldbraich, Ora Nova Fandina, Ateret Anaby-Tavor, Orna Raz, and Eitan Farchi. 2023. Unveiling safety vulnerabilities of large language models. *Preprint*, arXiv:2311.04124.

Lizhi Lin, Honglin Mu, Zenan Zhai, Minghan Wang, Yuxia Wang, Renxi Wang, Junjie Gao, Yixuan Zhang, Wanxiang Che, Timothy Baldwin, Xudong Han, and Haonan Li. 2024. Against the achilles' heel: A survey on red teaming for generative models. *Preprint*, arXiv:2404.00629.

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable bias: Analyzing societal representations in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

Moushumi Mahato, Avinash Kumar, Kartikey Singh, Bhavesh Kukreja, and Javaid Nabi. 2024. Red teaming for multimodal large language models: A survey.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *Preprint*, arXiv:2402.04249.

Microsoft. 2024. Python risk identification tool for generative ai (pyrit).

Manish Nagireddy, Lamogha Chiazor, Moninder Singh, and Ioana Baldini. 2024a. Socialstigmaqa: A benchmark to uncover stigma amplification in generative language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, (19):21454–21462.

Manish Nagireddy, Bernat Guillén Pegueroles, and Ioana Baldini. 2024b. Dare to diversify: Data driven and diverse llm red teaming. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6420–6421, New York, NY, USA. Association for Computing Machinery.

Manish Nagireddy, Inkit Padhi, Soumya Ghosh, and Prasanna Sattigeri. 2024c. When in doubt, cascade: Towards building efficient and capable guardrails. *Preprint*, arXiv:2407.06323.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*.

Jessica Quaye, Alicia Parrish, Oana Inel, Charvi Rastogi, Hannah Rose Kirk, Minsuk Kahng, Erin Van Liemt, Max Bartolo, Jess Tsang, Justin White, Nathan Clement, Rafael Mosquera, Juan Ciro, Vijay Janapa Reddi, and Lora Aroyo. 2024. Adversarial nibbler: An open red-teaming method for identifying diverse harms in text-to-image generation. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Bhaktipriya Radharapu, Kevin Robinson, Lora Aroyo, and Preethi Lahoti. 2023. AART: AI-assisted red-teaming with diverse data generation for new LLM-powered applications. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) Industry Track*.

Sander Schulhoff, Jeremy Pinto, Anaum Khan, Louis-François Bouchard, Chenglei Si, Svetlina Anati, Valen Tagliabue, Anson Kost, Christopher Carnahan, and Jordan Boyd-Graber. 2023. Ignore this title and HackAPrompt: Exposing systemic vulnerabilities of LLMs through a global prompt hacking competition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Rusheb Shah, Quentin Feuillade-Montixi, Soroush Pour, Arush Tagade, Stephen Casper, and Javier Rando. 2023. Scalable and transferable black-box jailbreaks for language models via persona modulation. *Preprint*, arXiv:2311.03348.

The Royal Society and Humane Intelligence. 2024. Red teaming large language models (LLMs) for resilience to scientific disinformation.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? In *Advances in Neural Information Processing Systems (NeurIPS)*.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical safety evaluation of generative ai systems. *Preprint*, arXiv:2310.11986.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

White House. 2023. Red-teaming large language models to identify novel ai risks.

Zheng Xin Yong, Cristina Menghini, and Stephen Bach. 2023. Low-resource languages jailbreak gpt-4. In *Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS*.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. 2024. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *Preprint*, arXiv:2401.06373.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.