Guardrails and Security for LLMs: Safe, Secure and Controllable Steering of LLM Applications

Traian Rebedea, Leon Derczynski, Shaona Ghosh, Makesh Narsimhan Sreedhar, Faeze Brahman, Liwei Jiang, Bo Li, Yulia Tsvetkov, Christopher Parisien, and Yejin Choi

Website: `https://llm-guardrails-security.github.io/`

Pretrained generative models, especially large language models, provide novel ways for users to interact with computers. While generative NLP research and applications had previously aimed at very domain-specific or task-specific solutions, current LLMs and applications (e.g. dialogue systems, agents) are versatile across many tasks and domains. Despite being trained to be helpful and aligned with human preferences (e.g., harmlessness), enforcing robust guardrails on LLMs remains a challenge. And, even when protected against rudimentary attacks, just like other complex software, LLMs can be vulnerable to attacks using sophisticated adversarial inputs. This tutorial provides a comprehensive overview of key guardrail mechanisms developed for LLMs, along with evaluation methodologies and a detailed security assessment protocol - including auto red-teaming of LLM-powered applications. Our aim is to move beyond the discussion of single prompt attacks and evaluation frameworks towards addressing how guardrailing can be done in complex dialogue systems that employ LLMs.

Traian Rebedea, Principal Research Scientist at NVIDIA and Associate Professor at University Politehnica of Bucharest.
Email: `trebedea@nvidia.com`
Website: `https://www.linkedin.com/in/trebedea/`
His research is focused mainly on dialogue and safety, on topics such as dialogue steering and improving multi-turn LLM safety and security. At the same time, he is an important contributor in developing Romanian models and datasets. He received his PhD from University Politehnica of Bucharest, Romania. Prior to joining NVIDIA, he co-founded Roboself and was Chief Data Scientist at Wholi, working on dialogue systems and information retrieval.

Leon Derczynski, Principal Research Scientist in LLM Security at NVIDIA, Associate Professor in NLP at ITU University of Copenhagen, & President of ACL SIGSEC.
Email: `lderczynski@nvidia.com`
Website: `https://www.linkedin.com/in/leon-derczynski/`
Prof Derczynski has organised many workshops and tasks in the past (multiple WNUT, multiple TempEval, multiple RumourEval, OffensEval), as well as co-chairing COLING 2018, ACing and SACing all the major ACL events, and EiCing a journal (NEJLT). He has held tutorials at NAACL 2023, COLING 2020, and EACL 2014.

Shaona Ghosh, Senior Research Scientist at NVIDIA.
Email: `shaonag@nvidia.com`
Website: `https://www.linkedin.com/in/shaonaghosh`
Dr Ghosh is Senior Research Scientist at NVIDIA, focusing on AI safety and leading efforts in LLM content moderation. She chairs the AI Risk and Reliability workstream at MLCommons, contributing to its global AI safety benchmark. She completed postdoctoral research at University of Cambridge and University of Oxford, and holds a PhD from the University of Southampton, in collaboration with UCL in the UK. Previously, she worked at Apple for six years on safety, robustness, and privacy in NLP, computer vision, and multimodal domains.

Makesh Narsimhan Sreedhar, Research Scientist at NVIDIA.
Email: `makeshn@nvidia.com`

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts), pages 13–15
July 27, 2025 ©2025 Association for Computational Linguistics

Website: https://www.linkedin.com/in/makeshsreedhar/

He is a Research Scientist at NVIDIA, working on AI Safety and model alignment techniques. His current research focuses on enhancing dialogue systems and improving the instruction-following capabilities of language models. He holds a Master's degree from the University of Wisconsin-Madison.

Faeze Brahman, Research Scientist at Allen Institute for AI.

Email: fae.brahman@gmail.com

Website: https://fabrahman.github.io/

She did her Ph.D. in Computer Science at the University of California, Santa Cruz. She is broadly interested in understanding language model's capabilities and limitations. More recently, she focused on AI alignment, trustworthy AI and robust evaluation of LLMs' safety in complex interactive tasks. She has organized multiple workshops at ACL and AAAI as well as AC'ing and SAC'ing major ACL conferences.

Liwei Jiang, PhD Student at the University of Washington and Graduate Research Intern at NVIDIA.

Email: lwjiang@cs.washington.edu

Website: https://liweijiang.github.io/

She is a Ph.D. candidate at Paul G. Allen School of Computer Science & Engineering, University of Washington, advised by Prof Yejin Choi. She is a graduate research intern at NVIDIA and was previously a student researcher at Allen Institute for AI. Her research centers on humanistic AI safety, currently focusing on pluralistic alignment, self-improving algorithms for steerable and secure language models, and anticipatory strategies for long-term risks, such as overreliance and the erosion of human creativity.

Bo Li, Associate Professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign.

Email: bol@uchicago.edu

Website: https://aisecure.github.io/

Prof Li is the recipient of several awards, including the IJCAI Computers and Thought Award, Alfred P. Sloan Research Fellowship, NSF CAREER Award, AI's 10 to Watch, MIT Technology Review TR-35 Award, and also best paper awards at several top machine learning and security conferences. Her research focuses on both theoretical and practical aspects of trustworthy machine learning, which is at the intersection of machine learning, security, privacy, and game theory. She has designed several scalable frameworks for robust learning and privacy-preserving data publishing systems.

Yulia Tsvetkov, Associate Professor at the University of Washington.

Email: yuliats@cs.washington.edu

Website: https://homes.cs.washington.edu/ yuliats/

Prof Tsvetkov is Associate Professor in the Paul G. Allen School of Computer Science & Engineering at the University of Washington. Her work is in natural language processing, with a focus on AI ethics and safety. Her lab develops models and algorithms to advance language technologies for high-stakes domains such as health, science, and education. This research integrates advances in machine learning with novel evaluation and alignment methods to ensure large language models serve diverse users and avoid harm.

Christopher Parisien, Senior Manager of Applied Research at NVIDIA.

Email: cparisien@nvidia.com

Website: https://www.linkedin.com/in/christopher-m-parisien/

Dr Parisien is leading research efforts in safety, security, and dialogue in large language models at NVIDIA. He holds a PhD in Computational Linguistics from the University of Toronto. He has served as a research scientist at Nuance, focused on virtual assistants and clinical language understanding, and as Chief Technology Officer at NexJ Health, a patient-centred health platform.

Yejin Choi, Professor and MacArthur Fellow at Stanford University, and Senior Director at NVIDIA.
Email: `yejinc@stanford.edu`
Website: `https://yejinc.github.io/`
Yejin Choi is Dieter Schwarz Foundation Professor of Computer Science and Senior Fellow at Human Centered Artificial Intelligence at Stanford University, Senior Director at NVIDIA, and an ACL Fellow. Prof Choi has presented multiple keynotes and won best and outstanding papers at related venues including NeurIPS and ACL, and held tutorials at ACL 2020, CVPR 2020, and COLING 2022.