IJCNLP-AACL 2023

The ART of Safety Workshop: Adversarial Testing and Red Teaming for Generative AI

Proceedings of the Workshop

November 1, 2023

The IJCNLP-AACL organizers gratefully acknowledge the support from the following sponsors.

Platinum

Gold

Silver

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-026-4

Message of thanks from the workshop organizers

We would like to thank the authors for their valuable workshop submissions, as well as the many anonymous reviewers who volunteered their time for this workshop. We also thank all the participants of the Adversarial Nibbler Challenge—your contributions will help us better understand safety in text-to-image models.

Organizing Committee

Table of Contents

Conference Program

Red Teaming for Large Language Models At Scale: Tackling Hallucinations on Mathematics Tasks
Aleksander Buszydlik, Karol Dobiczek, Michał Teodor Okoń, Konrad Skublicki, Philip Lippmann and Jie Yang

Student-Teacher Prompting for Red Teaming to Improve Guardrails
Rodrigo Revilla Llaca, Victoria Leskoschek, Vitor Costa Paiva, Cătălin Lupău, Philip Lippmann and Jie Yang

Distilling Adversarial Prompts from Safety Benchmarks: Report for the Adversarial Nibbler Challenge
Manuel Brack, Patrick Schramowski and Kristian Kersting

Measuring Adversarial Datasets
Yuanchen Bai, Raoyi Huang, Vijay Viswanathan, Tzu-Sheng Kuo and Tongshuang Wu

Discovering Safety Issues in Text-to-Image Models: Insights from Adversarial Nibbler Challenge
Gauri Sharma

Uncovering Bias in AI-Generated Images
Kimberley Baxter