

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

Nusa Dua, Indonesia
1 November 2023

ISBN: 978-1-7138-9003-4

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2023) by the Asian Federation of Natural Language Processing and the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2024)

For permission requests, please contact the Association for Computational Linguistics at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

| | |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| <i>Red Teaming for Large Language Models At Scale: Tackling Hallucinations on Mathematics Tasks</i> Aleksander Buszydlík, Karol Dobiczek, Michał Teodor Okoń, Konrad Skublicki, Philip Lippmann and Jie Yang | 1 |
| <i>Student-Teacher Prompting for Red Teaming to Improve Guardrails</i> Rodrigo Revilla Llaca, Victoria Leskoschek, Vitor Costa Paiva, Cătălin Lupău, Philip Lippmann and Jie Yang | 11 |
| <i>Distilling Adversarial Prompts from Safety Benchmarks: Report for the Adversarial Nibbler Challenge</i> Manuel Brack, Patrick Schramowski and Kristian Kersting | 24 |
| <i>Measuring Adversarial Datasets</i> Yuanchen Bai, Raoyi Huang, Vijay Viswanathan, Tzu-Sheng Kuo and Tongshuang Wu | 29 |
| <i>Discovering Safety Issues in Text-to-Image Models: Insights from Adversarial Nibbler Challenge</i> Gauri Sharma | 43 |
| <i>Uncovering Bias in AI-Generated Images</i> Kimberley Baxter | 49 |