ACL 2025

The First Workshop on LLM Security

Proceedings of the Workshop

August 1, 2025

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-279-4

Introduction

We are excited to present LLMSEC 2025, the first workshop of the ACL Special Interest Group in LLM and NLP Security, SIGSEC.

SIGSEC was proposed on 21 July, 2023, to to promote interest in security in natural language processing and security of natural language processing resources and techniques; to provide members of the ACL with a special interest in NLP security with a means of exchanging news of recent research developments and other matters of interest in the union of information security and computational linguistics; and to sponsor meetings and workshops in NLP Security that appear to be timely and worthwhile, operating within the framework of the ACL's general guidelines for SIGs.

The time leading up to this was a period of great change for NLP and for language modelling. Decades of research and many years of engineering excellent came together to bring language models to people in a way they could interact with efficiently and retrive results they found interesting or useful for everyday tasks, rather than mostly benchmarking, as was research tradition. With this contact between NLP and broad societal use came a cornucopia of expectations and dangers - including security risks. Attacks and defences quickly emerged.

As the fields of NLP, machine learning security, and traditional cybersecurity merge, we find that no one group has all the answers. We all need each other in order to make sense of the phenomena and interactions we observe. NLP researchers are not intrinsically experts in security; and cybersec experience brings no guarantees of learning any computational linguistics or machine learning. Further, this novel challenge cannot survive with just industrial or just academic input alone; industry sees and deflects advanced attacks rapidly without sharing details - academia uncovers new classes of techniques and deep analyses. This is a departure from traditional NLP research, but a dynamic commonly observed in security. We all have a lot to learn from each other.

And so the research LLMSEC includes the entire life cycle of LLMs, from training data through fine-tuning and alignment over to inference-time. It also covers deployment context of LLMs, including risk assessment, release decisions, and use of LLMs in agent-based systems.

For this, our first event, despite a compressed timeline we elicited 34 submissions, 16 of which were accepted on the basis of quality alone. Of these, six are presented as talks, and ten as posters. They are met by a set of excellent keynote talks from our speakers.

We are grateful to our highly diverse program committee, our paper authors, our speakers, and especially our audiuence, for their time and attention; we look forward to fruitful discussions and an exciting event.

Leon Dercyznski, Jekaterina Novikoa, Muhao Chen
Organizers, and chairs of ACL SIGSEC

Organizing Committee

Organizers

Leon Derczynski, NVIDIA Corporation / IT University of Copenhagen
Jekaterina Novikova, Vanguard
Muhao Chen, University of California, Davis

Program Committee

Program Committee

Mohammad Akbari, University College London
Hend Al-Khalifa, King Saud University
Nura Aljaafari, University of Manchester
Segun Taofeek Aroyehun, University of Konstanz
Lavish Bansal, AI Security Engineer
Farah Benamara, University of toulouse
Renaud Bidou, ParaCyberBellum
Adrian Brasoveanu, Modul University
Donato Capitella, Reversec
Canyu Chen, Illinois Institute of Technology
Yiyi Chen, Aalborg University
Olga Cherednichenko, freelancer
Pedro Cisneros-Velarde, VMware Research
Christina Dahn, GESIS
Alfonso De Gregorio, Zeronomicon
Ali Dehghantanha, University of Guelph
Chunrong Fang, Nanjing University
Elisabetta Fersini, University of Milano-Bicocca
Komal Florio, University of Torino (Italy) - HighEST Lab
Xingyu Fu, Upenn
Erick Galinkin, NVIDIA Corporation
Severi Giorgio, Microsoft
Anmol Goel, TU Darmstadt
Kerem Goksel, Independent Researcher
Rich Harang, NVIDIA Corporation
Yifeng He, University of California, Davis
Ales Horak, Masaryk University
Ken Huang, DistributedApps.ai
Umar Iqbal, Washington University in St. Louis
Jafar Isbarov, The George Washington University
Chao Jiang, Georgia Institute of Technology
Weizhao Jin, AWS
Akbar Karimi, University of Bonn
Sepehr Karimi Arpanahi, University of Tehran
Paritosh Katre, PayPal Inc
Gauri Kholkar, Pure Storage
Dan.klein Klein, Accenture
Valia Kordoni, Humboldt-Universität zu Berlin
Arjun Krishna, 8090
Hajar Lachheb, URV
Hwaran Lee, NAVER AI Lab
Heather Lent, Aalborg University
Jiazhao Li, Amazon.com
Tianhao Li, Duke University
Qin Liu, University of California, Davis
Yepang Liu, Southern University of Science and Technology

Yi Liu, Quantstamp
Dongqi Liu, Saarland University
Xingyu Lyu, University of Massachusetts Lowell
Viraaji M, Kennesaw State University
Danni Ma, University of Pennsylvania
Eugenio Martínez Cámara, University of Jaén
Stephen Meisenbacher, Technical University of Munich
George Mikros, Hamad Bin Khalifa University
Wenjie Mo, University of Southern California
María Navas-Loro, Universidad Politcnica de Madrid
Atul Kr. Ojha, Data Science Institute, Unit for Linguistic Data, University of Galway
Venkatesh Pala, Chase
Brian Pendleton, self
Ehsan Qasemi, Oracle
Yanjun Qi, University of Virginia
Changyuan Qiu, University of Michigan
Imranur Rahman, North Carolina State University
Tharindu Ranasinghe, Lancaster University
Elena Sofia Ruzzetti, University of Rome Tor Vergata
Martin Sablotny, NVIDIA Corporation
Anudeex Shetty, The University of Melbourne
Ankit Srivastava, OryxLabs
Adam Swanda, Cisco (Robust Intelligence)
Daniel Takabi, Old Dominion University
Liling Tan, Apple
Ali Tekeoglu, Leidos & JHU
S.m Towhidul Islam Tonmoy, Islamic University Of Technology
Fatih Turkmen, University of Groningen
Prasoon Varshney, NVIDIA Corporation
Haoyu Wang, Huazhong University of Science and Technology
Wayne Wang, University of Hong Kong
Fei Wang, University of Southern California
Jicheng Wang, University of California Davis
Adrian Wood, Dropbox
Tianyi Yan, University of Southern California
Yixiang Yao, Information Sciences Institute
Jingwei Yi, University of Science and Technology of China
Tianwei Zhang, Nanyang Technological University
Yubo Zhang, The Hong Kong Polytechnic University
Yang Zhong, University of Pittsburgh
Zining Zhu, Stevens Institute of Technology

Keynote Talk

A Bunch of Garbage and Hoping: LLMs, Agentic Security, and Where We Go From Here

Erick Galinkin
NVIDIA Corporation

Abstract: Large Language Models are, in some ways, a miracle. Despite a paucity of theoretical linguistic underpinning and a swath of known weaknesses, they have proven empirically successful beyond the wildest imaginings of many, leading to integration in a wide variety of applications. This has necessitated a strong response from both the information security community and those who study large language models.

This talk examines both cybersecurity implications of LLMs and the LLM implications of cybersecurity. We provide some background on adversarial examples in computer vision as a lens to view the problems in AI systems and cover the parlance of cybersecurity as it frames AI problems. Using these two lenses, we examine the state of LLM security and discuss approaches to uncover and mitigate the risks inherent in LLM-powered applications.

Bio: Erick Galinkin is a Research Scientist at NVIDIA working on the security assessment and protection of large language models. Previously, he led the AI research team at Rapid7 and has extensive experience working in the cybersecurity space. He is an alumnus of Johns Hopkins University and holds degrees in applied mathematics and computer science. Outside of his work, Erick is a lifelong student, currently at Drexel University and is renowned for his ability to be around equestrians.

Keynote Talk

What does it mean for agentic AI to preserve privacy?

Niloofar Mireshghallah
Meta/CMU

Bio: Dr. Mireshghallah is a Research Scientist at Meta AI's FAIR Alignment group and joins Carnegie Mellon University's Engineering Public Policy (EPP) Department and Language Technologies Institute (LTI) as an Assistant Professor in Fall 2026.
Her research interests are privacy, natural language processing, and the societal implications of ML. Dr. Mireshghallah explores the interplay between data, its influence on models, and the expectations of the people who regulate and use these models. Her work has been recognized by the NCWIT Collegiate Award and the Rising Star in Adversarial ML Award.

Keynote Talk
Linguistic Diversity in NLP Security

Johannes Bjerva
Aalborg University

Bio: Prof. Bjerva's research is characterised by an interdisciplinary perspective on NLP, with a focus on the potential for impact in society. His main contributions to my field are to incorporate linguistic information into NLP, including large language models (LLMs), and to improve the state of resource-poor languages. Recent research focuses on embedding inversion and attacks on multi-modal models.

Keynote Talk

Trust No AI - Prompt Injection Along the CIA Security Triad

Johann Rehberger
Independent

Abstract: The CIA security triad - Confidentiality, Integrity, and Availability - is a cornerstone of data and cybersecurity. With the emergence of large language model (LLM) applications, a new class of threat, known as prompt injection, was first identified in 2022. Since then, numerous real-world vulnerabilities and exploits have been documented in production LLM systems, including those from leading vendors like OpenAI, Microsoft, Anthropic and Google. This paper compiles real-world exploits and proof-of concept examples, based on the research conducted and publicly documented, demonstrating how prompt injection undermines the CIA triad and poses ongoing risks to cybersecurity and AI systems at large.

Furthermore the talk will explore command and control infrastructure for ChatGPT which is exploited entirely based on prompt injection and memory persistence.

Table of Contents

Program

Friday, August 1, 2025

09:00 - 09:05	*Opening Remarks*
09:05 - 09:55	*Erick Galinkin: A Bunch of Garbage and Hoping: LLMs, Agentic Security, and Where We Go From Here*
09:55 - 10:30	*Lightning and Posters 1*
11:00 - 10:30	*Break*
11:00 - 11:50	*Niloofar Mireshghallah: What does it mean for agentic AI to preserve privacy?*
11:50 - 12:50	*Papers 1*
14:00 - 12:50	*Lunch*
14:50 - 14:00	*Johannes Bjerva: Linguistic Diversity in NLP Security*
14:50 - 15:25	*Lightning and Posters 2*
16:00 - 15:25	*Break*
16:00 - 16:30	*Papers 2*
16:30 - 17:20	*Johann Rehberger: Trust No AI - Prompt Injection Along the CIA Security Triad*
17:20 - 17:25	*Best paper award, SIGSEC business, and closing*