Guest Editorial >>>

Special Topic on Security of Large Models

Guest Editors

SU Zhou

DU Linkang

Large models, such as large language models (LLMs), vision-language models (VLMs), and multimodal agents, have become key elements in artificial intelligence (AI) systems. Their rapid development has greatly improved perception, generation, and decision-making in various fields. However, their vast scale and complexity bring about new security challenges. Issues such as backdoor vulnerabilities during training, jailbreaking in multimodal reasoning, and data provenance and copyright auditing have made security a critical focus for both academia and industry.

This special issue on the security of large models aims to highlight recent advances in uncovering, analyzing, and addressing critical vulnerabilities that arise in the age of large models. The five selected papers represent a diverse and timely exploration of attack methodologies, defense mechanisms, and system-level frameworks that are reshaping our understanding of trustworthy AI.

The first paper, titled "Poison-Only and Targeted Backdoor Attack Against Visual Object Tracking", reveals a novel poison-only backdoor threat in the context of visual object tracking (VOT). The authors propose a targeted attack strategy that manipulates full video sequences via a method called Random Video Poisoning (RVP), exploiting temporal correlations to inject stealthy backdoors. This paper identifies three invariant categories based on size, trajectory, and hybrid alterations. The authors demonstrate that these attacks can precisely manipulate object tracking while preserving high performance on unaltered data. Beyond security implications, the findings also suggest possible extensions to watermarking and forensic tracing.

The second paper, "VOTI: Jailbreaking Vision-Language Models via Visual Obfuscation and Task Induction", investigates the vulnerability of VLMs to multimodal jailbreak attacks. The authors design a two-pronged attack strategy that subtly hides malicious queries in visual inputs through visual obfuscation and decomposes harmful prompts into subtasks via task induction. This approach successfully bypasses global safety mechanisms in multiple mainstream VLMs, achieving a 73.46% attack success rate on GPT-4o-mini. The work sheds light on over-reliance on local visual cues and lack of robust multi-step reasoning alignment, calling for stronger cross-modal defenses.

The third paper, "From Function Calls to MCPs for Securing AI Agent Systems: Architecture, Challenges and Countermeasures", provides a comprehensive study of the Model Context Protocol (MCP), a new integration interface for LLM-based agents proposed by Anthropic. As MCP gains traction in orchestrating tool use and environmental interactions, its security remains underexplored. This paper presents the first systematic analysis of MCP's architectural vulnerabilities, demonstrates a real-world tool-injection attack, and categorizes mitigation strategies. The authors conclude with forward-looking insights for securing MCP-powered AI agent systems under realistic adversarial conditions.

The fourth paper, "Dataset Copyright Auditing for Large Models: Fundamentals, Open Problems, and Future Directions", provides a structured survey on the emerging field of dataset copyright auditing for large model training. The paper categorizes existing auditing methods by data modality, train-

DOI: 10.12142/ZTECOM.202503001
Citation (Format 1): SU Z, DU L K. Editorial: security of large models [J]. ZTE Communications, 2025, 23(3): 1–2. DOI: 10.12142/ZTECOM.202503001
Citation (Format 2): Z. Su and L. K. Du, "Editorial: security of large models," *ZTE Communications*, vol. 23, no. 3, pp. 1–2, Sept. 2025. doi: 10.12142/ZTECOM.202503001.

ing stage, data overlap, and model access. It identifies key trends such as the dominance of black-box auditing and the limited focus on pre-training. The authors review 12 pivotal works and propose future research directions. They stress the importance of a standard benchmark for thoroughly comparing current methods, improving robustness at low watermark rates, and proposing auditing strategies for multimodal datasets.

The fifth paper, "StegoAgent: A Generative Steganography Framework Based on GUI Agents", introduces a novel steganographic approach that embeds hidden information in the behavioral traces of GUI agents, such as mouse movements and clicks, rather than altering traditional carriers. Using LLM-based agents and an entropy-adaptive encoding scheme, StegoAgent achieves high-capacity, accurate, and stealthy information hiding in both offline and real-time interactive environments, demonstrating agent trajectories as an effective new covert communication channel.

To conclude, this special issue provides an in-depth exploration of large model security from multiple perspectives: input manipulation, agent protocol integrity, training data verification, and novel covert channels. We hope that these articles will inspire future research on the trustworthy AI systems.

We express our sincere appreciation to all the authors for their excellent contributions, to the reviewers for their professional insights and timely feedback, and to the editorial team for their dedicated support throughout this process.

Biographies

SU Zhou is a professor with Xi'an Jiaotong University, China, and his research interests include multimedia communication, wireless communication, network security and network traffic. Dr. SU has published technical papers in top journals and conferences, including *IEEE JSAC*, *IEEE/ACM ToN*, *IEEE TWC*, and IEEE INFOCOM. He received the Best Paper Awards at international conferences including IEEE AIoT 2024, IEEE WCNC 2023, IEEE VTC-Fall 2023, and IEEE ICC 2020. He is an Associate Editor of *IEEE Internet of Things Journal* and *IEEE Open Journal of Computer Society*. He is also the chair of the IEEE VTS Xi'an Section Chapter.

DU Linkang received his BE and PhD degrees from Zhejiang University in 2018 and 2023, respectively. He is currently an assistant professor at the School of Cyber Science and Engineering, Xi'an Jiaotong University, China. He has published technical papers in top security conferences, including IEEE Symposium on Security and Privacy, USENIX Security, NDSS, and ACM CCS. His research interests include trustworthy machine learning and privacy-preserving computing.