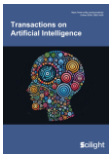
Scilight

Transactions on Artificial Intelligence

https://www.sciltp.com/journals/tai



Editorial

Perspective on Artificial Intelligence for Security

Cong Wang [1] and Wei Bao [2,*]

[1] Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
[2] School of Computer Science, University of Sydney, Sydney, NSW 2006, Australia
* Correspondence: wei.bao@sydney.edu.au

How To Cite: Wang, C.; Bao, W. Perspectives on Artificial Intelligence for Security. *Transactions on Artificial Intelligence* **2025**, *1*(1), 197–198. https://doi.org/10.53941/tai.2025.100012.

The unprecedented rise of Artificial Intelligence (AI) in recent years has not only revolutionized automation, reasoning, and language understanding but has also introduced new vectors of risk and vulnerability. As AI systems, especially large language models (LLMs), become increasingly embedded in safety-critical domains such as healthcare, finance, and infrastructure, their security becomes not just a technical concern but also a societal imperative. This special issue on AI Security in the Transactions on Artificial Intelligence (TAI) brings together contemporary research on the threats, defenses, and emerging paradigms associated with secure AI deployment.

This collection of works provides both depth and breadth, spanning areas such as backdoor vulnerabilities in LLMs, embedding-as-a-service attacks, secure program analysis, and federated learning in privacy-sensitive domains like medical imaging. The issue reflects the rapidly evolving attack surface of modern AI systems and the pressing need for robust and explainable defenses.

The paper titled "A Survey on Backdoor Threats in Large Language Models (LLMs): Attacks, Defenses, and Evaluation Methods" [1] offers a comprehensive overview of the growing threat of training-time backdoor attacks. As LLMs increasingly power downstream applications through APIs and third-party tools, their susceptibility to stealthy manipulations during training has significant implications. The paper classifies current attack mechanisms, outlines state-of-the-art defense strategies, and highlights open evaluation challenges, laying a foundation for future research in adversarially robust LLM development.

Closely related, the paper titled "Is Embedding-as-a-Service Safe? Meta-Prompt-Based Backdoor Attacks for User-Specific Trigger Migration" [2] investigates a novel threat model targeting EaaS (Embedding-as-a-Service), a paradigm where users remotely access LLM embeddings for downstream customization. The proposed attack method, BadEmd, introduces meta-prompt-based backdoors and adaptive trigger migration techniques that retain stealthiness while achieving cross-user attack efficacy. This work presents a concrete example of how modular AI services, when insufficiently safeguarded, can introduce systemic risks.

Stepping back from attack vectors and focusing on positive use cases, the paper titled "A Contemporary Survey of Large Language Model Assisted Program Analysis" [3] reviews how LLMs are reshaping traditional static, dynamic, and hybrid code analysis methods. Program analysis plays a vital role in software security, and the infusion of LLM-based reasoning holds immense potential—but also raises concerns regarding hallucinations, false positives, and explainability. This survey helps delineate the capabilities and limitations of LLMs in code understanding, offering a valuable roadmap for researchers seeking to bridge language models and secure software engineering.

Lastly, the paper titled "Federated Learning for Medical Image Analysis: Privacy-preserving Paradigms and Clinical Challenges" [4] addresses a critical frontier of AI security—privacy. In healthcare, data sharing is limited by regulatory constraints, yet collaborative learning remains essential for improving clinical outcomes. This paper systematically categorizes federated learning (FL) techniques, not only in terms of privacy and security, but also regarding architectural strategies and unlearning techniques. It underscores the complexity of balancing privacy, utility, and domain-specific constraints, particularly in sensitive medical imaging applications.

Collectively, the papers in this collection present a timely, multi-faceted view of AI security. From adversarial manipulation to privacy-aware learning, the issue highlights the dual mandate of advancing AI performance while

ensuring robustness and accountability. The emphasis on both attack taxonomies and defense strategies serves as a foundational reference for the community.

We extend our sincere appreciation to all authors for their thoughtful contributions and to the reviewers for their rigorous feedback. We also thank the editorial team at TAI for supporting this important collection. As AI systems continue to scale, their security must evolve in tandem, and we hope it serves as a springboard for further innovation at the intersection of AI and trustworthy computing.

Collection Editors:

Cong Wang (Fellow, IEEE) is currently a Professor with the Department of Computer Science, City University of Hong Kong. His research interests include data and network security, blockchain and decentralized applications, and privacy-enhancing technologies. He is a member of the ACM. He has been the Founding Members of the Young Academy of Sciences of Hong Kong since 2017, and has been conferred the RGC Research Fellow in 2021. He was a recipient of the Outstanding Researcher Award (Junior Faculty) in 2019, the Outstanding Supervisor Award in 2017, the President's Awards in 2019 and 2016, all from City University of Hong Kong; a co-recipient of the Best Paper Award of IEEE ICDCS 2020, ICPADS 2018, MSN 2015; and the Best Student Paper Award of IEEE ICDCS 2017, and IEEE INFOCOM Test of Time Paper Award 2020. His research has been supported by multiple government research fund agencies, including the National Natural Science Foundation of China, Hong Kong Research Grants Council, and Hong Kong Innovation and Technology Commission. He was an Associate Editor for IEEE Transactions on Dependable and Secure Computing, IEEE Transactions on Services Computing, IEEE Internet of Things Journal, IEEE Networking Letters, and The Journal of Blockchain Research, and the TPC co-chair for a number of IEEE conferences and workshops.

Wei Bao (Member, IEEE) received the BE degree in communications engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2009, the MASc degree in electrical and computer engineering from the University of British Columbia, Vancouver, Canada, in 2011, and the PhD degree in electrical and computer engineering from the University of Toronto, Toronto, Canada, in 2016. He is currently an Associate Professor with the School of Computer Science, the University of Sydney, Sydney, Australia. His research covers the area of network science, with particular emphasis on Internet of things, mobile computing, edge computing, and distributed learning. He received the Best Paper Awards in ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM) in 2013 and 2019 and IEEE International Symposium on Network Computing and Applications (NCA), in 2016.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Zhou, Y.; Ni, T.; Lee, W.-B.; et al. A Survey on Backdoor Threats in Large Language Models (LLMs): Attacks, Defenses, and Evaluation Methods. *Trans. Artif. Intell.* **2025**, *1*, 28–58.
2. Bagwe, G.; Zhang, L.; Guo, L.; et al. Is Embedding-as-a-Service Safe? Meta-Prompt-Based Backdoor Attacks for User-Specific Trigger Migration. *Trans. Artif. Intell.* **2025**, *1*, 16–27.
3. Wang, J.; Ni, T.; Lee, W.-B.; et al. A Contemporary Survey of Large Language Model Assisted Program Analysis. *Trans. Artif. Intell*. **2025**, *1*, 105–129.
4. Hu, J.; Yang, Z.; Wang, P.; et al. Federated Learning for Medical Image Analysis: Privacy-Preserving Paradigms and Clinical Challenges. *Trans. Artif. Intell*. **2025**, *1*, 153–169.