Special Session 7: Security and Privacy for Large Language Models (LLMs)

Chairs: Guang Chen, Beijing University of Posts and Telecommunications, China

Brief Description of the Session

Large Language Models (LLMs) are transforming the technological landscape with their impressive capabilities in natural language processing, code generation, and more. This rapid advancement, however, introduces significant security and privacy risks that demand immediate attention. This track aims to foster discussion and collaboration on cutting-edge research addressing these crucial challenges. We invite contributions exploring the vulnerabilities of LLMs, including adversarial attacks like prompt injection, backdoors, and data poisoning, as well as defenses against these threats. Furthermore, we encourage submissions focusing on privacy-preserving techniques for LLM training and inference, such as federated learning, differential privacy, and homomorphic encryption, to safeguard sensitive data. Explainability and transparency are also key concerns, and we welcome research on methods for interpreting LLM decision-making and promoting accountability. Finally, this track will delve into the broader ethical implications of LLMs, covering bias detection and mitigation, fairness, and responsible AI development. Join us to contribute to the development of secure, privacy-respecting, and ethically sound LLMs, paving the way for their responsible and beneficial deployment in society.

Topics

- **Adversarial Attacks and Defenses:** Exploring vulnerabilities of LLMs to adversarial attacks (e.g., prompt injection, data poisoning) and developing robust defense mechanisms.
- **Privacy-Preserving Training and Inference:** Investigating techniques like federated learning, differential privacy, and homomorphic encryption to protect sensitive data used in training and inference.
- **Explainability and Transparency:** Developing methods to understand and interpret LLM decision-making processes, enhancing transparency and accountability.
- **Data Security and Integrity:** Addressing issues related to data breaches, model theft, and ensuring the integrity of training data and model outputs.
- **Ethical Considerations and Responsible AI:** Examining the ethical implications of LLM deployment, including bias detection and mitigation, fairness, and accountability.
- **Secure LLM Deployment and Operations:** Exploring secure infrastructure and practices for deploying and managing LLMs in real-world applications.
- **Legal and Regulatory Frameworks:** Discussing the legal and regulatory landscape surrounding LLM security and privacy and proposing guidelines for responsible development and use.

Brief Introducation of Chair and Co-chairs with Photo

Guang Chen is an Associate Professor at the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. He has participated in multiple national-level projects, including the National Natural Science Foundation of China, as a key researcher. He has published numerous papers in leading journals and international conferences and authored a textbook. His research interests include Machine Learning and Language AI.