# Special Session on Security and Trustworthiness of Large AI Models and Agents (SecTrust-AI)

Recent advances in large-scale artificial intelligence models have revolutionized modern AI, delivering powerful capabilities in natural language understanding, logical reasoning, code generation, computer vision, multimodal perception, speech processing, and scientific discovery. Large language models (LLMs), vision-language models (VLMs), and general generative AI models have greatly advanced the deployment of intelligent systems across diverse real-world scenarios. Built on these foundational models, AI agents have evolved into autonomous entities that support complex planning, tool invocation, memory management, environmental perception, and collaborative decision-making. Such agentic AI systems exhibit promising applicability in software engineering, scientific research, healthcare, financial services, intelligent robotics, and cyber-physical systems.

Nevertheless, the widespread deployment of large AI models and autonomous agents brings unprecedented security, safety, privacy, and trustworthiness risks, which have become critical obstacles to reliable AI adoption. Large models are inherently vulnerable to prompt injection, jailbreak exploitation, adversarial manipulation, model stealing, data poisoning, and private information leakage. Compared with standalone models, AI agents further broaden the attack surface due to their autonomous planning logic, external tool interactions, long-term memory mechanisms, multi-agent collaboration, and continuous dynamic environment perception. These emerging threats make the security and trustworthiness of large-model and agentic AI systems a vital research topic for cybersecurity and trustworthy AI communities.

Held within the 11th IEEE Cyber Science and Technology Congress, this special session aims to gather academic and industrial researchers to report state-of-the-art research on the security, privacy, robustness, and trustworthiness of large AI models and intelligent agents. We welcome original research papers covering novel theories,

technical methodologies, optimization algorithms, system designs, practical applications, and benchmark evaluations that advance the secure, reliable, and accountable deployment of large-model and agent-based AI systems.

Prospective authors are invited to submit original manuscripts via the conference submission portal at https://edas.info/, and please select the special session titled "Security and Trustworthiness of Large AI Models and Agents" during submission. All accepted papers will be published in IEEE conference proceedings and indexed by IEEE Xplore and EI Compendex. High-quality accepted papers will be recommended for journal special issue publication after substantial extension and rigorous revision. For updated conference policies and arrangements, please refer to the official website: https://cyber-science.org/2026/cyberscitech/.

Topics include, but are not limited to:

- Security and Privacy of Large AI Models
- AI Agent Security and Trustworthiness
- Prompt Injection and Jailbreak Attacks
- Adversarial Attacks and Robustness of Large AI Models
- Security of Multimodal AI Models
- Secure Tool Use and Function Calling
- Multi-Agent System Security
- Agent Memory Security and Information Leakage
- Retrieval-Augmented Generation (RAG) Security
- Model Alignment, Safety, and Guardrails
- AI Red Teaming and Automated Vulnerability Discovery
- Privacy-Preserving Training and Inference for Large AI Models
- Federated and Distributed Large Model Security
- Model Watermarking, Fingerprinting, and Intellectual Property Protection
- Detection and Mitigation of Hallucinations and Unsafe Behaviors
- Trustworthy Evaluation and Benchmarking for Large AI Models and AI Agents
- AI Safety for Autonomous Decision-Making Systems

- Applications of Secure Large AI Models and AI Agents in Cybersecurity
- Large AI Models and AI Agents for Cyber Defense and Threat Intelligence
- Responsible, Explainable, and Ethical AI Systems

**Chair:**

Yukuan Yang, Institute of Software, Chinese Academy of Sciences, China, yangyukuan@iscas.ac.cn

**Co-Chairs：**

Yueming Wu, Huazhong University of Science and Technology, China, yuemingwu@hust.edu.cn

Chao Lin, Nanjing University of Aeronautics and Astronautics, China, chaolin@nuaa.edu.cn

**Committee:**

Jiasi Weng, Guangzhou University, China, wengjiasi@gzhu.edu.cn

Man Yao, Institute of Automation, Chinese Academy of Sciences, China, man.yao@ia.ac.cn

Yihan Lin, Xiamen University, China, linyh@xmu.edu.cn

**Important Dates:**

Paper Submission Due: August 15, 2026

Acceptance Notification Due: September 11, 2026

Final Manuscript Due: October 1, 2026