Jiaxiong Tang

jxtang1@stu.ecnu.edu.cn
github.com/dickton

EDUCATION

East China Normal University 2024.09 - 2027.06
Software Engineering Master Shanghai, China
- GPA: 3.29 / 4.0

Research Interests: Trustworthy AI, Privacy & Ownership issues in AI

Dalian University of Technology 2020.09 - 2024.06
Cyber Engineering Bachelor Dalian, China
- GPA: 3.35 / 5.0

SKILLS

- Professional Skills: Architecture design, model selection, benchmarking, and ablation-driven iteration; Python, Pytorch; C++; Agentic workflow building; familiar with LLM fine-tuning & inference.
- Languages: Chinese (Native), English (Proficiency, IELTS 6.5), Japanese (JLPT N2)

RESEARCH EXPERIENCE

Robust Client-Server Watermarking for Split Federated Learning 2025.02 - 2025.06
First Author
- Designed and implemented a server–client joint watermarking framework to address model leakage and ambiguous ownership attribution in Split Federated Learning (SFL), enabling verifiable model ownership proof under the SFL architecture.
- Experimental results showed that the method maintained strong main-task performance, achieved over 95% detection accuracy with statistical significance ($p<0.03$), and preserved high watermark retention under attack scenarios.
- Outperformed baseline methods and showed good compatibility with large-scale SFL across multiple model architectures.
- Led code development, experimental design and evaluation, and paper writing, with deep involvement throughout the full research lifecycle from problem formulation to paper output.

Sigil: Server-Enforced Watermarking in U-Shaped Split Federated Learning via Gradient Injection 2025.06 - 2025.11
Co-author
- Given the limited model access in U-SFL, the Intellectual Property Rights of the server is under threat. Thus proposed an stealthy server-to-client watermarking embedded via gradient injection.
- Sigil achieves high detection rates, rubosness to SOTA anomalty detection methods and attacks in SFL.
- Highly engaged in project development, experiment evaluating and paper drafting. The paper is on the ***Arxiv***.

Self-adaptive Differential Privacy Framework for Split Federated Learning 2025.06 - 2026.03
- Addressed private data reconstruction risks in SFL, where existing dynamic differential privacy methods typically depend on full model knowledge or strong expert priors, leading to high deployment costs.
- Proposed an adaptive privacy-risk indicator using only local client-side information and developed a dynamic privacy budget allocation algorithm for client-limited settings, with its effectiveness validated through statistical experiments.
- Demonstrated through multiple SOTA privacy inference attacks that the method improved privacy protection by about 10% over expert-dependent baselines while preventing reconstruction of privacy data and maintaining comparable task accuracy.
- Led the project from problem formulation to paper output, covering method design, implementation, evaluation, and writing.

Stateful Backdoor Attacks on LLM Agents 2026.02 - Present
- Proposed a stateful backdoor attack framework leveraging Agent Memory to enable stealthy, cross-session attack coordination in LLM agents.
- Fine-tuned Qwen3.5 and LLaMA3 with QLoRA in a simulated supply-chain poisoning scenario, allowing persistent attack states and covert malicious behaviors during authorized tool use.
- Validated the approach across multiple open-source LLMs, achieving high attack success rates despite existing safety alignment.
- Led method design, experimental studies, and paper writing across the full research workflow.

PROJECT EXPERIENCE

Small Target Detection in Extreme Noise via U-Net DNN 2025.08 - 2025.09
Technical Lead Collaborated with the National Space Science Center, CAS
- Led the end-to-end project lifecycle, independently handling requirements communication, algorithm development, system implementation, delivery testing, and iterative refinements.
- Developed an environmental simulation algorithm and implemented a U-Net network with 3D-convolution.
- Achieved 95%+ PD, <0.4% PFA and <1.0 px ALE under SNR=2dB (rough environment) while traditional method is invalid.
- Delivered the full system under a tight 6-week timeline with successful expert acceptance.