Zhengchunmin Dai

Software Engineering Institute, East China Normal University | Shanghai, China
51275902069@stu.ecnu.edu.cn | Homepage | Google Scholar | ORCID

Education

M.Sc. **Software Engineering**, East China Normal University — Sep 2024 – present

B.Sc. **Software Engineering**, East China Normal University — Sep 2020 – Jul 2024

Research Appointments

Research Assistant — Oct 2026 – Present
The Hong Kong University of Science and Technology (Guangzhou)
Research focus: LLM security and mechanistic interpretability.

Research Interests

LLM and agent security from a sub-model perspective (see My Backdoor Agenda). Currently interested in the intersection of security, mechanistic interpretability, and alignment, with a practical focus. Example questions:

- White- and black-box spoofing of control-token representations and its security implications.
- Constructing and evaluating model organisms of hidden scheming in latent reasoning models.
- The control geometry of refusal and its apparent asymmetry with acceptance.

Publications

[1] **Stateful Agent Backdoors: Constructing Cross-Session Attack Programs.**
Zhengchunmin Dai, Jiaxiong Tang, Liantao Wu, Zhibo Wang, Peng Sun, Honglong Chen, Yihao Huang, Zhe Zhao.
Preprint; under review, arXiv:2605.06158v2 (2026). *Lead author.*

[2] **Server-Enforced Watermarking in U-Shaped Split Federated Learning.**
Zhengchunmin Dai, Jiaxiong Tang, Peng Sun, Honglong Chen, Liantao Wu.
Preprint, arXiv:2511.14422v2 (2025; revised 2026). *Lead author.*

[3] **Robust Client-Server Watermarking for Split Federated Learning.**
Jiaxiong Tang, **Zhengchunmin Dai**, Liantao Wu, Peng Sun, Honglong Chen.
IEEE Internet of Things Journal, 13(17): 40185–40197, 2026. *Co-first author.*
doi:10.1109/JIOT.2026.3709448

Research Projects

Steering Prompt Injection with Prefix Label Directions
Research report | Blog on LessWrong Oct 2026 – Nov 2026

- Proposed a prefix method that adds preceding empty role blocks to neutral-text inputs for training role probes; evaluated probing and steering across three models and two datasets.
- The prefix method improved toolness probing over the original method and enabled PI steering in some settings. Effects varied substantially across models and datasets, showing mismatches between role readouts and intervention effects.

Stateful Agent Backdoors: Constructing Cross-Session Attack Programs
Lead author | arXiv:2605.06158v2 Jan 2026 – Sep 2026

- Independently led problem formulation, method design and formalization, training-data generation, implementation, experiments, and main drafting.
- Represented cross-session attack programs as Mealy machines. Constructed single-session training trajectories for individual state transitions and jointly fine-tuned a model to learn these local behaviors; persistent memory connects them into a complete program after a single initial trigger injection.
- Constructed and evaluated four program instantiations covering sequential execution with conditional waiting, branching and merging, looping, and condition accumulation.
- Evaluated four models in LangChain: the primary instantiation achieved mean complete-program success rates of 71.7–94.7%. In the official OpenClaw runtime under a controlled configuration, the primary instantiation achieved 85% for each of two evaluated models.

Server-Enforced Watermarking in U-Shaped Split Federated Learning
Lead author | arXiv:2511.14422v2 Sep 2024 – Dec 2025

- Independently led the project end-to-end. Enabled a split-learning server to embed an ownership watermark in client-side models through the gradient channel, without access to client data, labels, or model parameters.
- Defined a watermark constraint in server-visible activations and used adaptive gradient clipping to limit task interference and reduce detectability. Designed an adaptive removal attack and evaluated robustness against removal attacks.

Robust Client-Server Watermarking for Split Federated Learning
Co-first author | *IEEE Internet of Things Journal*, 2026 Sep 2024 – Dec 2025

- Co-designed a dual-side scheme protecting both client- and server-side model ownership; contributed to method formulation and writing.

Skills

Python, PyTorch; LLM inference and supervised fine-tuning (LoRA/QLoRA, PEFT); agent pipelines and evaluation (LangChain, OpenClaw); federated and split learning.

Languages

Mandarin Chinese (native), English (research language), Japanese (conversational).