Yiwen Tu

✉ y2tu@ucsd.edu
☎ 734-(877)-0677

Research Interest

My research interests broadly revolve around building **reliable** and **trustworthy** machine learning models, particularly in the context of **large language models (LLMs)**. Specifically, I am interested in **privacy** and **interpretability**.

Education

- **University of California, San Diego** — **United States**
  Master of Science in Computer Science, GPA: 3.91/4.00 — *2024.09-2026.06*
  - **Selected Courses**: ML System(A+), Search and Optimization(A), Differential Privacy(A), ML for Music(A), Computer Security(A-), Computer Vision (In Progress)
- **University of Michigan, Ann Arbor** — **United States**
  *Bachelor in Computer Science with **Summa Cum Laude**, GPA: 3.94/4.00* — *2022.09-2024.05*
  - **Selected Courses**: CV(A), NLP(A), Machine Learning Theory(A+), Convex Optimization(A), Database Management Systems(A), Cryptography(A)
- **Shanghai Jiao Tong University** — **China**
  Bachelor of Electrical Computer Engineering (Dual Degree), GPA: 3.67/4.00 — *2024.09-2026.06*
  - **Selected Courses**: Mathematical Analysis(A+), Differential Equations(A+), Discrete Mathematics(A+), Linear Algebra(A+)

Peer-Reviewed Conference Publications

(* denotes equal contribution)

[C1] Tu, Yiwen*, Hu, Pingbang*, Ma, Jiaqi., "A Reliable Cryptographic Framework for Empirical Machine Unlearning Evaluation.". In *Proceedings of the 39th Advances in Neural Information Processing Systems* (NeurIPS 2025)

[C2] Tu, Yiwen*, Liu, Ziqi*, Tang, Weijing, Ma, Jiaqi., "Measuring Fine-Grained Relatedness in Multitask Learning via Data Attribution.". In *2nd Attributing Model Behavior at Scale (ATTRIB) Workshop at 38th Advances in Neural Information Processing Systems* (NeurIPS 2024 ATTRIB Workshop)

[C3] Ma, Jiaqi*, Zhang, Xingjian*, Fan, Hezheng, Huang, Jin, Li, Tianyue, Li, Ting Wei, Tu, Yiwen, Zhu, Chenshu, Mei, Qiaozhu., "Graph Learning Indexer: A Contributor-Friendly and Metadata-Rich Platform for Graph Learning Benchmarks.". In *Proceedings of the First Learning on Graphs Conference* (LOG 2022 **Oral**)

Research Experience

- **DATASMITH Lab, University of California, San Diego** — **California, USA**
  Researcher supervised by Prof. Haojian Jin and Prof. Lianhui Qin — *Jul 2025 – Present*
  - **Individual-level Privacy Concerns Reasoning**: Proposed an agent architecture that bridges existing privacy and cognitive theories and individual-level reasoning on privacy concerns. The agent structure reconstructs user-specific "privacy minds" and dynamically activates context-relevant beliefs, achieving substantial gains over naive concept bottleneck models.
  - In submission to a top-tier conference in natural language processing.
- **Trustworthy AI Lab, University of California, San Diego** — **California, USA**
  Researcher supervised by Prof. Lily Weng — *Jan 2025 – Present*
  - **Fine-Grained Concept Bottlenecks Large Language Models**: Enhanced concept-bottleneck large language models through LLM synthetic data augmentation, LLM-driven concept labeling, multi-label steering mechanisms, concept-steering training loss, and hierarchical bottleneck designs, yielding stronger intrinsic interpretability and controllable generation in tasks like controlled text generation and question answering.
  - In preparation for a top-tier conference in machine learning.
- **TRAIS Lab, University of Illinois Urbana–Champaign** — **Illinois, USA**
  Researcher supervised by Prof. Jiaqi Ma — *Jul 2024 – May 2025*
  - **Instance-level Multitask Influence Framework**: Developed the first scalable instance-level influence-function framework for multitask learning, enabling precise identification and diagnosis of positive and negative transfer on a per-instance basis.
  - Accepted by **NeurIPS 2024 ATTRIB Workshop**.
- **TRAIS Lab, University of Illinois Urbana–Champaign** — **Illinois, USA**
  Researcher supervised by Prof. Jiaqi Ma — *May 2023 – Oct 2024*

    - **Machine Unlearning Evaluation Framework**: Introduced a cryptography-inspired metric to quantify residual data leakage in approximate data-deletion scenarios, supported by rigorous theoretical analyses and empirical validation.
    - Accepted by **NeurIPS 2025**.

- **FOREST Lab, University of Michigan, Ann Arbor** — **Michigan, USA**
  Research Assistant supervised by Prof. Qiaozhu Mei — *May 2022 – Oct 2022*
    - **Graph Neural Network Benchmark Platform**: Developed a scalable, contributor-friendly platform for graph learning benchmarks, optimizing usability and metadata management across datasets with millions of nodes and edges.
    - Accepted by **LoG 2022 Oral**.

Teaching Experience

- **Grader, University of Michigan** — **Ann Arbor, USA**
  Course Grader — *Sep 2023 – Dec 2023*
    - **EECS 487: Introduction to NLP**: Graded weekly assignments on language modeling, seq2seq translation, and transformer architectures for a cohort of ˜120 students.
- **Grader, University of Michigan** — **Ann Arbor, USA**
  Course Grader — *Jan 2024 – Apr 2024*
    - **EECS 484: Database Management System**: Graded weekly assignments on database systems for a cohort of ˜120 students.
- **Instructional Aide, Shanghai Jiao Tong University** — **Shanghai, China**
  Instructional Aide — *Sept 2021 – Aug 2022*
    - **Mathematical Analysis I & II**: Facilitated weekly discussion and office hours for a 90+ student analysis class in English, answered questions, clarified materials, and led review sessions.

Honors and Awards

- **Undergraduate Excellence Scholarship** — **Shanghai, China**
  Shanghai Jiao Tong University — *2021*
- **Student Development Scholarship** — **Shanghai, China**
  Shanghai Jiao Tong University — *2021*
- **Dean's List** — **Ann Arbor, USA**
  University of Michigan — *2022–2024*
- **James B. Angell Scholar** — **Ann Arbor, USA**
  University of Michigan — *2024*
- **Finalist, Mathematical Contest in Modeling** — **Global**
  Recognized as a finalist team in the international Mathematical Contest in Modeling. — *2022*

Professional Service

- **Program Committee**
  AAAI 2026
- **Conference Reviewer**
  ICML 2024, NeurIPS ATTRIB Workshop 2024, ICLR 2025, AISTATS 2026