Yen-Shan (Lily) Chen

(+886) 902-339-190 yenshan.ntu@gmail.com www.linkedin.com/in/yen-shan-lily-chen

Publications

[1] **Yen-Shan Chen***, Zhi Rui Tam*, Cheng-Kuang Wu, and Yun-Nung Chen. "Expected Harm: Rethinking Safety Evaluation in (Mis)Aligned LLMs." Under Review at *Conference on Neural Information Processing Systems (NeurIPS)*. 2026. [paper link]

[2] **Yen-Shan Chen**, Sian-Yao Huang, Cheng-Lin Yang, and Yun-Nung Chen. "TraceSafe: A Systematic Assessment of Trajectory-Level Safety in Tool-Calling Agents." In *Conference on Language Modeling (COLM)*. 2026. [paper link]

[3] **Yen-Shan Chen***, Shih-Yu Lai*, Ying-Jung Tsou, Yi-Cheng Lin, Bing-Yu Chen, Yun-Nung Chen, Hung-yi Lee, and Shang-Tse Chen. "Latent-Mark: An Audio Watermark Robust to Neural Resynthesis." In *Annual Conference of the International Speech Communication Association, INTERSPEECH*. 2026. [paper link]

[4] **Yen-Shan Chen**, Sian-Yao Huang, Cheng-Lin Yang, and Yun-Nung Chen. "Eyes-on-Me: Scalable RAG Poisoning through Transferable Attention-Steering Attractors." In *International Conference on Machine Learning (ICML)*. 2026. [paper link]

[5] **Yen-Shan Chen***, Jing Jin*, Peng-Ting Kuo, Chao-Wei Huang, and Yun-Nung Chen. "LLMs are Biased Evaluators but Not Biased for Fact-Centric Retrieval Augmented Generation." In *Findings of the Association for Computational Linguistics: ACL*. 2025. [paper link]

Selected Technical Talks

[1] **Yen-Shan Chen** and Cheng-Lin Yang. "One Poisoned Artifact Can Steer Your AI: How Robust Are Your LLM-Assisted Security Workflows?" In *FIRST CON*. Denver, USA, June 2026.

[2] **Yen-Shan Chen**. "One Bad OSINT Can Ruin Everything: How Secure is Your CTI RAG System?" In *FIRST CTI*. Munich, Germany, April 2026. [video]

[3] **Yen-Shan Chen**, Sian-Yao Huang, and Cheng-Lin Yang. "BullyRAG: A Multi-Perspective RAG Robustness Evaluation Framework" In *Code Blue*. Tokyo, Japan, Oct. 2024.

[4] **Yen-Shan Chen**, Sian-Yao Huang, and Cheng-Lin Yang. "LLMs Bear the BOT, Yet Swallow Up: Attacks Against RAG Models." In *SinCon*. Singapore, Aug. 2024.

Work Experience / Lab Affiliations

CyCraft Technology Corporation — March 2023 – Present

Data Scientist (Data Scientist Intern before Sep. 2025) — *New Taipei City, Taiwan*

- Leading the technical design and benchmarking of **TraceSafe**, a trajectory-level safety evaluation suite for multi-step tool-calling agents; engineered a testing pipeline for 20+ proprietary and open-source guardrails.
- Developed **Eyes-on-Me**, a modular RAG poisoning framework utilizing attention-steering attractors; optimized transferable attacks across 10+ LLM architectures. (**ICML 2026**)
- Engineered an automated robustness suite (**BullyRAG**) to detect vulnerabilities in RAG workflows; implemented word- and paragraph-level obfuscation techniques to stress-test security-centric LLM pipelines (**Code Blue**, **SinCon 2024**).
- Integrated research findings into production-grade cyber threat intelligence insights, presenting technical deep-dives to international audiences of 1000+ security professionals at **FIRST CON** and **FIRST CTI**.

Machine Intelligence and Understanding Lab, NTU — Sep. 2023 – Present

Graduate Researcher (Undergraduate Researcher before Sep. 2025; Supervised by Prof. Yun-Nung Chen) — *Taipei, Taiwan*

- Led the design of evaluation frameworks for **self-preferential bias** in RAG systems and **probabilistic safety** metrics; developed the core codebase for reranking and generation phase experiments (**Findings of ACL 2025**).
- Benchmarked **Expected Harm** by combining proposed metrics with SOTA jailbreak methods (e.g., GCG, AutoDAN) across 4 major safety datasets; performed mechanistic study through linear probing and staged checkpoint analysis.
- Supervising two undergraduate research groups, providing technical mentorship on **multilingual fairness** in retrieval and the intersection of **Theory of Mind (ToM)** with LLM behavioral alignment.
- Proposed the theoretical intuition of **latent invariance** for robust audio watermarking; engineered a framework to embed signals within a neural codec's latent space to withstand semantic neural compression.

California Institute of Technology — June 2024 – August 2024

Summer Research Intern (Supervised by Prof. Colin Camerer) — *California, USA*

- Investigated the mechanisms of **Rational Inattention** by bridging information costs with cognitive workload; analyzed high-frequency eye-tracking data to validate behavioral economic models.

- Engineered data pipelines to process **multidimensional gaze trajectories**, extracting temporal features from pupil dilation and fixation patterns to quantify cognitive effort.
- Developed experimental interfaces using **PsychoPy** and applied **unsupervised natural language clustering** to categorize subject responses for multiple lab projects, mapping behavioral strategies to latent decision-making patterns.

Google **June 2023 – August 2023**

Software Engineering Intern (Pixel Camera Team) *Taipei, Taiwan*

- Engineered **subject detection and spatial reasoning** algorithms for the Google Pixel Camera to optimize group photo framing and human-centric identification.
- Developed **clustering heuristics** and integrated ML detectors to improve target subject isolation in complex multi-person scenes; work contributed to **Guided-Frame** feature launch featured in the **2024 Super Bowl** advertisement.

National Taiwan University **Sep. 2022 – Present**

(Lead) Teaching Assistant (AI & Deep Learning Courses) *Taipei, Taiwan*

- **Foundations of Artificial Intelligence (Spring 2026):** Led the engineering of a multi-agent game engine for a tournament final project; developed baseline heuristic/RL agents and automated evaluation infrastructure.
- **Applied Deep Learning (Fall 2025):** Engineered a two-stage **adversarial jailbreak challenge** for 150+ students, requiring the design of both external guardrails and intrinsic model safety filters.
- Awarded **Excellent Teaching Assistant Certificate** (Top 10% university-wide) for mentoring 300+ students across Calculus and Algorithm Design (2022-2023); formulated complex problem sets and algorithm optimization challenges.

Awards & Honors

- **Garmin Student Scholarship** (2025): One of 14 STEM graduate students selected nationwide in Taiwan.
- **E.SUN Bank Graduate Student Scholarship** (2025): Awarded for excellence in AI/Technology research.
- **ACLCLP Student Travel Grant** (2025): Awarded by the Association for Computational Linguistics and Chinese Language Processing to support international conference attendance.
- **ICPC Asia Taoyuan Regional**: Silver Award & Best Female Team Award (2023).
- **Presidential Award**, National Taiwan University (2×): Awarded to the top 5% of the class.
- **National First Prize**, Girls in CyberSecurity Contest (2022): Research on smart healthcare vulnerabilities.
- **Fu Bell Scholarship** (2021): Award of academic excellence (awarded to 1 student per dept).

Professional Service & Leadership

Academic Service | *Peer Review* **Jan. 2026 – Present**

- **Official Reviewer:** COLM 2026, NeurIPS 2026; **Sub-Reviewer:** ICML 2026.

Taiwan AI Safety Club | *Technical Discussion* **Aug. 2025 – Present**

- Contribute to weekly deep-dive discussions on AI alignment, robustness, and mechanistic interpretability; synthesize international safety research to facilitate local knowledge exchange.

Ambassador, Girls in CyberSecurity | *National Outreach Campaign* **Sep. 2023 – Present**

- Invited to mentor participants on adversarial ML and security research; advocate for gender diversity in STEM through technical guidance and career mentorship.

Education

National Taiwan University **Sep. 2025 – Present**

M.S. in Computer Science and Information Engineering, GPA 4.3/4.3 *Taipei, Taiwan*

National Taiwan University **Sep. 2021 – June 2025**

B.S. in Computer Science and Information Engineering, Double Major in Economics, GPA: 4.12/4.3 *Taipei, Taiwan*

- Thesis: Understanding Self-Preferential Bias in Large Language Models (Advisor: Prof. Yun-Nung Chen)
- Honors: **Outstanding Bachelor's Thesis Award** (Top 2 among 170+ students)

Skills

- **Languages: English** (IELTS Academic: **8.0** | R:8.5, L:9.0, S:8.0, W:7.0), **Mandarin** (Native)
- **Programming & Frameworks:** Python (Expert), C/C++, Shell, PyTorch, Hugging Face (Transformers, PEFT), LangChain, Linux, LaTeX.
- **AI Safety & Security:** Adversarial Robustness, RAG Poisoning, Jailbreak Benchmarking, and LLM Guardrail evaluation/training.