Pura Peetathawatchai
(607) 379-4035 | pura@stanford.edu | poonpura.github.io

INTERESTS

AI security and privacy: privacy attacks against machine learning models, adversarial machine learning, differential privacy, machine unlearning, cybersecurity risk posed by LLM agents

Other interests: cryptography, AI for climate change

EDUCATION

Stanford University **Stanford, California**
M.S. Computer Science (Specialization in Cybersecurity and Artificial Intelligence) *March 2025 (expected)*

- Cumulative GPA: 3.91 / 4.00
- Relevant Coursework: Deep Learning, Deep Generative Models, Cryptography, Cybersecurity, Bioinformatics, Computer Networks, Trust and Safety

Cornell University **Ithaca, New York**
B.A. (Dist.) Mathematics and Computer Science (double major) *May 2022*

- Cumulative GPA: 4.04 / 4.00
- Honors: Member of Phi Beta Kappa Honors Society (junior inductee - top 3% of graduating class), graduated with distinction in all subjects
- Relevant Coursework: Cryptography, Machine Learning, Natural Language Processing, Computer Architecture, Operating Systems, Design and Analysis of Algorithms, Mathematical Logic, Chaos Theory, Abstract Algebra, Number Theory, Real Analysis, Decision Theory, Game Theory, Cognitive Science

EXPERIENCE

Stanford University **Stanford, California**
*Graduate Researcher (under Profs. **Dan Boneh, Percy Liang, Daniel Ho**)* *Apr 2024 - present*

- Developed Cybench: a benchmark for evaluating cybersecurity capabilities of LLMs, which includes 40 professional-level Capture-the-Flag (CTF) tasks chosen to be recent, meaningful and spanning wide range of difficulties
- Reviewed, edited and standardized subtasks, which break down CTF task into intermediary steps to enable more granular evaluation, across 17 CTF challenges
- Led team to annotate CTF tasks with meaningful metadata for incorporation into LLM agent and environment
- Comparing top human hackers against their AI counterparts to analyze differences in behavior and performance gap

Stanford Trustworthy AI Research **Stanford, California**
*Graduate Researcher (under Prof. **Sanmi Koyejo**)* *Jan 2024 - present*

- Proposed and investigating use of Textual Inversion and DreamBooth as new method of privacy-preserving style transfer for diffusion models (with Prof. **Albert No**, paper in progress)
- Implementing and assessing various approaches to incorporate differential privacy into Universal Guidance for diffusion models and Textual Inversion
- Investigated susceptibility of LoRA-trained transformer models to membership inference attacks in relation to fully fine-tuned models
- Co-authored paper accepted to *Conference on Language Modeling (COLM)* 2024 (28.8% acceptance rate)
- Implemented Likelihood Ratio Attack (LiRA), a state-of-the-art membership inference attack in Python

Siametrics Consulting **Bangkok, Thailand**
AI Engineer Intern *Jun 2024 - Sept 2024*

- Prototyped LLM-powered assistants using LangChain and Streamlit with capabilities for document search/summarization (via Retrieval-Augmented Generation) and SQL querying for Stock Exchange of Thailand and Tourism Authority of Thailand
- Addressed and assured clients regarding data and infrastructure security concerns
- Delivered company-wide presentation on AI security and privacy concepts (slides: https://tinyurl.com/aispsmc)

Stanford Machine Learning Group **Stanford, California**

*Graduate Researcher (under Prof. **Andrew Ng**)* *Jan 2023 - Jun 2023*

- Engineered neural networks (DenseNet, Swin Transformer) to attribute methane plumes to landfills and concentrated animal feeding operations (CAFOs) via satellite imagery
- Implemented additional functionality to codebase enabling streamlined support for multispectral input, semantic segmentation and granular error analysis
- Investigated temporal and geographical distribution shifts of landfill satellite imagery and robustness of deep learning models to such distribution shifts
- Examined potential improvements to training procedure to reduce false positives, including two-tiered approach involving passing input through two separately trained models during inference

Cornell University **Ithaca, New York**

*Undergraduate Researcher (under Prof. **Noah Stephens-Davidowitz**)* *Jun 2021 – Dec 2021*

- Implemented many efficient lattice algorithms discussed in recent literature, including lattice sieving and discrete Gaussian basis sampling, in Python.
- Investigated extent to which LLL and BKZ reduction algorithms solve SVP on integer lattice basis vectors
- Analyzed effectiveness and weaknesses of different sampling techniques in generating secure bases as public keys

PUBLICATIONS

1. **Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risk of Language Models**
   Andy Zhang, Neil Perry, Riya Dulepet, Eliot Jones, Justin Lin, Joey Ji, Celeste Menders, Gashon Hussein, Samantha Liu, Donovan Jasper, **Pura Peetathawatchai**, Ari Glenn, Vikram Sivashankar, Daniel Zamoshchin, Leo Glikbarg, Derek Askaryar, Mike Yang, Teddy Zhang, Rishi Alluri, Nathan Tran, Rinnara Sangpisit, Polycarpos Yiorkadjis, Kenny Osele, Gautham Raghupathi, Dan Boneh, Daniel E. Ho, Percy Liang
   Preprint, 2024
2. **On Fairness of Low-Rank Adaptation of Large Models**
   Zhoujie Ding, Ken Ziyu Liu, **Pura Peetathawatchai**, Berivan Isik, Sanmi Koyejo
   Conference on Language Modeling (COLM), 2024
3. **Just how hard are rotations of Z^n? Algorithms and cryptography with the simplest lattice**
   Huck Bennett, Atul Ganju, **Pura Peetathawatchai**, Noah Stephens-Davidowitz
   International Conference on the Theory and Applications of Cryptographic Techniques (Eurocrypt), 2023

TEACHING

Cornell University *(Teaching Assistant)* **Ithaca, New York**

CS 4830/5830: Introduction to Cryptography *Spring 2022*
CS 2802: Discrete Structures - Honors *Spring 2020*
CS 1110: Introduction to Computing with Python *Fall 2019*

SKILLS

Programming Languages: Python (proficient), Java (familiar), OCaml (familiar), JavaScript (familiar), C/C++ (familiar)

Other Skills: PyTorch, Scikit-Learn, TensorFlow, NumPy, Pandas, SQL, LaTeX, HuggingFace, Stable Diffusion, Opacus, LangChain, Streamlit