Dzung V. Pham

PhD Candidate in Computer Science, University of Massachusetts Amherst
Mobile: +1 (413) 629-9759 | *E-mail:* dzungpham@cs.umass.edu
Personal Website | Google Scholar | GitHub | LinkedIn

Research Interests

I work on the security and privacy of LLMs and AI agents. I develop attacks and benchmarks that show how these systems leak data, can be manipulated through prompt injection, or act against their users, and I design defenses that are practical to deploy.

Keywords
Areas: security, privacy, safety
Systems: machine learning, large language models, AI agents, federated learning, the web
Topics: agent security, prompt injection, adversarial ML, differential privacy, leakage detection, evaluation

Education

University of Massachusetts Amherst **2022–Present**
PhD Candidate in Computer Science (GPA: 4.00). Expected graduation: 2028.
Advisors: Prof. Amir Houmansadr and Prof. Eugene Bagdasarian.

Williams College **2016–2020**
B.A. in Computer Science & in Statistics with Honors. GPA: 3.90 (Magna Cum Laude).
Thesis: *Exploring Variable Importance with Stacked Models*. Advisor: Prof. Richard D. De Veaux.

Publications

Can Prompt Anonymity Protect Your Identity From LLM Providers? Dzung Pham, Dillon Sheils, Naina Singh, Amir Houmansadr. Preprint 2026. [code]

AI Snitches Get Glitches: Towards Evading Agentic Surveillance. Hyejun Jeong,* Dzung Pham,* Amir Houmansadr, Eugene Bagdasarian. Preprint 2026. (*Equal contribution) [code]

AgentStop: Terminating Local AI Agents Early to Save Energy in Consumer Devices. Dzung Pham, Kleomenis Katevas, Ali Shahin Shamsabadi, Hamed Haddadi. ACM CAIS 2026. [code]

Can Large Language Models Really Recognize Your Name? Dzung Pham, Peter Kairouz, Niloofar Mireshghallah, Eugene Bagdasarian, Chau Minh Pham, Amir Houmansadr. ACM FAccT 2026. [code]

Frankentext: Stitching Random Text Fragments into Long-form Narratives. Chau Minh Pham, Jenna Russell, Dzung Pham, Mohit Iyyer. ACL 2026.

ProxyGPT: Enabling User Anonymity in LLM Chatbots via (Un)Trustworthy Volunteer Proxies. Dzung Pham, Jade Sheffey, Chau Minh Pham, Amir Houmansadr. NDSS 2025 Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb). [code]

RAIFLE: Reconstruction Attacks on Interaction-based Federated Learning with Adversarial Data Manipulation. Dzung Pham, Shreyas Kulkarni, Amir Houmansadr. NDSS 2025. [code]

Evaluating ProDirect Manipulation in Hour of Code. Quan Do, Kiersten Campbell, Emmie Hine, Dzung Pham, Alex Taylor, Iris Howley, Daniel Barowy. ACM SPLASH-E 2019.

Research Experience

AI Security Lab at UMass Amherst **2022–Present**

Research on secure and private AI systems with Prof. Amir Houmansadr and Prof. Eugene Bagdasarian. Led four (co)first-author projects on the security and privacy of LLMs and AI agents:

- **PromptAnonBench** (Preprint 2026): Built a benchmark of real human–LLM conversations to empirically evaluate the risk of prompt authorship re-identification when identity anonymity is applied. Found that at 10% false acceptance rate of out-of-set users, a nearest neighbor attack on historical conversations' embeddings can correctly re-identify 50–75% of SWE-Chat users and 10% of WildChat users.
- **SurveilBench** (Preprint 2026): Built a 330-scenario benchmark that tests whether AI agents report their users to third parties. Found that 6 of 10 frontier LLMs do so unprompted; a system prompt optimized on one model raises reporting to 75–97% on 9 of 10. Developed three prompt-injection evasion techniques that cut reporting below 10% on 7 of 10 models or misdirect 99% of reports.
- **AmBench** (ACM FAccT 2026): Built a benchmark of >12,000 real but ambiguous human names to stress-test LLM-based PII detection. Showed that 12 state-of-the-art LLMs detect these names with 20–40% lower recall than common names, and that benign prompt injections can quadruple name leakage in Clio, Anthropic's privacy-preserving conversation analysis pipeline.
- **ProxyGPT** (MADWeb 2025): Designed and built a peer-to-peer, Tor-backed anonymity layer for LLM chatbots that uses zkTLS (e.g., TLSNotary) to verify the integrity of volunteer proxies.
- **RAIFLE** (NDSS 2025): Developed a client-data reconstruction attack against federated recommendation and learning-to-rank systems that injects adversarial noise into the training features. Outperforms prior attacks by a wide margin, reaching >0.8 ROC AUC on popular recommendation datasets.
- **In progress**: A privacy-conscious orchestration pipeline for multi-agent systems.

Research Scientist Intern | Brave Software, Inc. **Summer 2025**

Developed **AgentStop** (ACM CAIS 2026), a lightweight runtime supervisor that monitors low-cost execution signals of a local AI agent, such as token-level log-probabilities, predicts trajectories unlikely to succeed, and terminates them early. Cuts wasted energy by up to 20% with $< 5\%$ task utility loss on agentic question-answering and coding benchmarks (SimpleQA, FRAMES, and SWE-Bench Verified). Built a measurement system for profiling local agents' power consumption on Apple Silicon and Nvidia Jetson devices. Mentors: Kleomenis (Minos) Katevas, Ali Shahin Shamsabadi.

Undergraduate Research | Williams College **2018–2020**

- **Senior Thesis**: Evaluated feature importance's robustness in stacked models; found that ensemble diversity can moderate the elevated importance of correlated features. Advisor: Prof. Richard De Veaux.
- **FallNet**: Developed a two-stream CNN to detect people falling in videos using TensorFlow and OpenCV. Won the 2019 Ward Prize for Best Project in Computer Science. Advisor: Prof. Duane Bailey.
- **SWELL**: Designed and implemented major parts of the SWELL graphical programming language, and taught it to 5th-grade students at Williamstown Elementary School. Advisor: Prof. Dan Barowy.

Industry Experience

Meta Platforms (Facebook), Inc.

Software Engineer – Machine Learning **2020–2022**

- Trained and deployed production ML models (e.g., GBDT, regression, neural nets) that detect and prevent fraud, scams, and harassment on Facebook Marketplace.
- Built and maintained scalable, reliable ML backend infrastructure that supports the integrity of Marketplace across billions of daily interactions.

Software Engineer Intern **Summer 2018 & 2019**

- Built two internal tools to support software release engineers with managing version releases and debugging software build failures.
- Trained and deployed a ranking ML model to help Facebook employees discover internal job opportunities based on their career history, skills and preferences.

Skills

Programming Languages: Python, Java, JavaScript, C/C++, SQL, Haskell, PHP
Frameworks & Tools: PyTorch, SLURM, Docker, vLLM/llama.cpp, HuggingFace Transformers
ML Methods: Large Language Models, Federated Learning, Adversarial Machine Learning, Membership Inference & Reconstruction Attacks, LLM Fine-tuning and Post-training (e.g., via RLHF, PPO)

Honors & Awards

The Internet Society NDSS Symposium Fellowship **2025**

Sigma Xi, Phi Beta Kappa, Mu Sigma Rho | Williams College **2020**
Elected to Sigma Xi (national honorary scientific research society), Phi Beta Kappa (academic honor society), and Mu Sigma Rho (national honorary society for statistics).

Ward Prize for Best Student Project in Computer Science | Williams College **2019**
Won the prize for the FallNet research project. Also nominated for the 2020 Ward Prize for a project on building a distributed reinforcement learning system.

Invited Talks

Can Large Language Models *Really* Recognize Your Name? **2025**
Google's Contextual Integrity Reading Group, UMass COMPSCI 690F Fall 2025
Guest Lecture | UMass COMPSCI 360: Intro to Computer Security **2025**
Towards Anonymous Chatbots with (Un)Trustworthy Browser Proxies
Guest Lecture | UMass First Year Seminar: Exploring Modern Computing **2024**
Large Language Models and User Privacy
Computer Science Colloquium | Williams College **2023**
On the Challenges of Building Privacy-preserving Chatbot Services

Teaching & Mentorship Experience

Teaching Associate | UMass Amherst **Fall 2025, Spring 2026**
Instructor of Record for COMPSCI 110: Foundations of Programming. Taught fundamental concepts of programming with Python to 80+ undergraduate Computer Science majors and non-majors. Reshaped the course's examination structure to de-emphasize rote memorization by introducing the use of "cheat sheet" and to promote learning from mistakes by pioneering an in-person exam revision mechanism.

PhD Mentor for Undergraduate Research Volunteer Program | UMass Amherst Summer 2026
Mentored two undergraduate students in conducting novel Computer Science research in evaluating the privacy guarantees of anonymous LLM proxy services. Paper under submission to ICLR 2027.

PhD Mentor | UMass Center for Data Science **2024–2025**
Mentored UMass Computer Science Master's students in collaboration with industry researchers from Google Research and Meta's FAIR Lab on evaluating the privacy properties of zeroth-order machine learning algorithms via membership inference attacks.

Organizer for COMPSCI 692PA: Seminar on AI Security and Privacy | UMass Amherst 2025
Helped program a graduate-level seminar on GenAI security and privacy with Prof. Eugene Bagdasarian at UMass. Invited and hosted leading experts to give talks. `https://aisec.cs.umass.edu/seminar`.

Graduate Teaching Assistant | UMass Amherst **2022–2024, Fall 2026**
Courses: COMPSCI 684: Trustworthy and Responsible AI, COMPSCI 360: Intro to Computer and Network Security, COMPSCI 326: Web Programming, COMPSCI 220: Programming Methodology.

Teaching Assistant | Williams College **2017–2020**
Courses: Statistical Learning & Data Mining, Regression & Forecasting, Principles of Programming Languages, Algorithm Design & Analysis, Computer Organization, Data Structures, Intro to Computer Science.

Service

Paper Reviewer: NeurIPS, ICLR, USENIX Security, The WebConf, IMC, IEEE TDSC.

PhD Applicant Support Program and Application Review | UMass Amherst **2024**
Supported applicants from under-represented communities in their applications to UMass CS PhD program (e.g., resume and statement of purpose review, one-on-one counseling). Reviewed Security/ML System applications to UMass CS PhD program.

Computer Science Student Advisory Committee | Williams College **2018–2019**
Organized the first mock tech interview program in the Computer Science department at Williams College to help fellow students apply for software engineering jobs in the tech industry.

Personal Projects

CSLAC: Computer Science at Liberal Arts Colleges **2026**
Creator of the cslac.org website (open-sourced on GitHub), which showcases data on 700+ Computer Science faculty and 1000+ courses at 130+ liberal arts colleges (LACs) in the US. CSLAC aggregates faculty data from each LAC's department and combines with Google Scholar citation metrics to shed light on the landscape of Computer Science education and research in LACs.