

Tai D. Nguyen

🌐 taidnguyen.github.io ✉ tainguyen7597@gmail.com ☎ (425) 246-3778

Summary	I currently work on post-training (multimodal/multilingual) models powering Apple Intelligence features. I previously worked on research for data attribution.		
Education	University of Pennsylvania	Philadelphia, PA	
	Master of Science in Engineering, Computer and Information Science	2021-2024	
	Thesis: Influence-based data attribution for in-context learning		
	Haverford College	Haverford, PA	
	Bachelor of Arts, Economics & Statistics	2015-2019	
	Thesis: Quantify the impact of home sharing on rental and welfare in New York City		
Experience	Apple	Cupertino, CA	
	Research Engineer	Oct 2024 - Present	
	- Lead training for: Multimodal autofill for web/app/PDFs, MessageSmartReply (watchOS), CalendarEventCreation (iOS), Proofreading (iOS) - Scale post-training recipes across languages and modality - Create billion-token scale synthetic data pipelines for training and evaluation		
	Allen Institute for AI (AI2)	Remote	
	Research Assistant	March 2024 - Oct 2024	
	- Advisor: Ben Bogin - Developed DataDecide framework for optimal pretraining data selection using small-scale experiments - Published at ICML 2025 main conference, with oral award at DataWorld Workshop		
	IBM	Burlington, VT	
	Data Scientist	2019 - 2021	
	- Built ML-powered analysis tools supporting 1200+ engineers developing IBM System-z servers - Productized algorithms that identified 100+ defective machine components, preventing costly failures - Won company-wide hackathon for new hires via a fraud detection Android app		
Selected Publications	DataDecide: How to Predict Best Pretraining Data with Small Experiments		
	I. Magnusson*, N. Tai*, B. Bogin*, D. Heineman, J. Hwang, L. Soldaini, A. Bhagia, J. Liu, D. Groeneveld, O. Tafjord, N. Smith, P. W. Koh, J. Dodge ICML 2025, DataWorld Workshop (Oral)		
	Massive Multilingual Text Embedding Benchmark		
	K. Enevoldsen, I. Chung, ... N. Tai ..., N. Muennighoff (82 authors) ICLR 2025		
	In-context example selection with influences		
	N. Tai, E. Wong Arxiv 2024		
	Explanation-based finetuning makes models more robust to spurious cues		
	J. Ludan, Y. Meng*, T. Nguyen*, S. Shah*, Q. Lyu, M. Apidianaki, C. Callison-Burch ACL 2023		

Software entity recognition with noise-robust learning
T. Nguyen, Y. Di, J. Lee, M. Chen, T. Zhang
ASE 2023

Ongoing Work	Terminal Bench 3.0	2026
	Contribute difficult security tasks on LLM inference stack fingerprinting.	
	Terminal Bench Science	2026
	Contribute tasks in gynecology & obstetrics with domain expert.	
Projects	NFL Big Data Bowl	2022
	<i>Finalist - \$15,000 Award</i>	
	- Top 5 finalist among 200+ submissions in a major football analytics competition	
	- Designed regression model simulating optimal kick return paths and new evaluation metrics	
	Chris Callison-Burch Lab	Philadelphia, PA
	<i>Student Lead</i>	May 2023 - Feb 2024
	- Led team of 3 on IARPA HIATUS project for interpretable authorship attribution	
	- Developed fully interpretable system of lexical features that outperform neural embeddings on cross-domain evaluation	
	Underthesea NLP	2021 - 2022
	<i>Contributor</i>	
	- Helped maintain a highly popular Vietnamese open-source toolkit - Improved Vietnamese NER and sentiment classification models by 15% F1 score	
Technical Skills	Programming: Python, C++, C, Bash, JavaScript	
	Framework: PyTorch, JAX, NumPy, Pandas	
	Languages: Vietnamese (native), English (native), Mandarin (intermediate)	
Honors & Awards	Vietnam Education Foundation (VEF) Fellow	2021
	IBM Volunteer Excellence Award	2021
	IBM New Developer Jumpstart Winner	2020

Last updated June 16, 2026.