# Lexington Whalen

https://www.linkedin.com/in/lxaw/

## SUMMARY


- Machine Learning Research Leadership: Co-Lead the development of flagship diffusion language models at NVIDIA, designing novel training and inference infrastructure for models of 1B to 8B parameters. Work was presented directly to CEO Jensen Huang and training / inference recipes used by 5+ internal research teams. Designs led to ~10x higher speed and ~10% improvements in accuracy as compared to leading models at the time.
- Led interdisciplinary teams of Ph.D. and Masters students in pioneering diffusion model research, achieving 5x training efficiency improvements while preserving model quality. Experience in efficient training and inference strategies, publishing at premier computer vision conferences, with research methodologies adopted by NVIDIA for internal development. Demonstrated expertise in optimizing training and inference strategies for both state-space and transformer architectures.
- Machine Learning Engineering: Developed comprehensive evaluation and training pipelines for NVIDIA's internal diffusion-LLM projects, scaling model training from millions to billions of parameters across billions to trillions of tokens. Experience with distributed training systems leveraging thousands of GPUs with advanced parallelization strategies.
- Technical Lead on United States' THOR healthcare initiative, developing innovative tiny PyTorch models that reduced cancer detection sensor degradation by over 70%. Collaborate with researchers at top institutions like Carnegie Mellon University, Stanford, Purdue, Georgia Tech, et cetera. Present to funding representatives, assist in developing grant reports.
- Recognition: NSF Graduate Research Fellowship recipient ($150,000 value), University of South Carolina Top Scholar Award ($100,000+ value), and Critical Language Scholar of Japanese (U.S. Department of State, only 500 selected among 5000+ applicants).
- Technical Development: Full-stack developer for research study websites using Django/React/SQL, managing databases with thousands of participant entries. Developed and won grants totaling $20,000+, leading teams of developers.


## EXPERIENCE

**Mar. 2025 – Present** **NVIDIA (Atlanta, Georgia & Santa Clara, California)**
**Efficient Deep Learning Intern**

- Train, finetune, evaluate the flagship series of NV Research 1-8B diffusion language models. Built the evaluation and inference infrastructure for diffusion language models; used internally by NV Research (5+ research teams).
- Worked on model distillation, efficient training recipes, efficient inference recipes. Built training and evaluation infra for base and instruct finetuned models.

- Specialize in diffusion models on high-performance computing infrastructure utilizing clusters of thousands of NVIDIA GPUs on SLURM-managed servers.
- Present weekly progress reports and technical findings to audiences of 20+ senior researchers and cross-functional engineering teams.
- Developed EfficientDLM series of diffusion language models that can outperform Qwen3 autoregressive series in both accuracy and speed. Over 500k internal downloads.
- Work with vision teams to release Nemotron series of diffusion language models; expected release in March of 2026.

**Dec. 2024 – Dec 2025** **NVIDIA (Atlanta, Georgia)**
**Data Filtering Challenge Lead**

- Led development and management of challenge website utilizing Google Analytics to track engagement and optimize user experience across 30+ participating university research teams and companies.
- Assisted in profiling of initial 400M parameter baseline model for participants, creating standardized benchmarks to evaluate submission quality.
- Designed and implemented a comprehensive evaluation framework leveraging 10B token fine-tuning datasets to ensure consistent, fair assessment of model submissions. Managed the evaluation process for participant submissions, establishing standardized performance benchmarks.
- Promoted challenge at International Conference on Machine Learning, ranked in the top 3 machine learning conferences globally with ~9,000 attendees and <25% paper acceptance rate.
- Coordinated with corporate sponsors including Lambda Labs and Turing to secure GPU resources, evaluating team submissions and managing the end-to-end competition workflow.

**Aug. 2024 – Dec 2025** **Georgia Institute of Technology (Atlanta, Georgia)**
**Machine Learning Engineer**

- Led team of 3 Ph.D. and Masters students to design innovative methods of reducing diffusion model training time by up to 5x on average while maintaining generation quality. Method was accepted to the top conference on computer vision in the world (Computer Vision and Pattern Recognition 2025)
- Algorithm lead for United States' ARPA-H THOR healthcare initiative. Spearheaded effort to reduce cancer detection sensor degradation by over 70%, improving the lifetime of the sensors from only a few hours to several days (over a 10x improvement)
- Engineered resource-efficient machine learning architectures designed to be powered by energy from the human body.
- Collaborated with 5+ teams from top universities like Stanford, Northwestern, Carnegie Mellon, and MIT. I represent Georgia Tech.
- Travel nationally to present technical progress to federal sponsors and healthcare stakeholders, effectively communicating complex technical concepts to diverse audiences.

**Jan. 2021 – Aug. 2024** **University of South Carolina (Columbia, South Carolina)**
**Software Developer**

- Led the design and implementation of three full-stack research study websites using PHP/Django/React/SQL, enabling efficient data collection and analysis for research studies.
- Architected and managed SQL databases of thousands of participant entries, ensuring data integrity, security compliance, and optimized query performance.
- Led the development of competitive federal and state grants applications, winning $20,000+ in funding.
- Led teams of 4+ developers, providing mentorship and technical guidance while meeting KPIs and performance targets.
- Implemented advanced machine learning techniques including clustering algorithms, random forests, and neural networks to analyze patient data, identify patterns, and develop predictive models that enhanced research outcomes and clinical insights.
- Developed an innovative automated language similarity analysis system using numpy and pandas that reduced document comparison time by approximately 99% (from 2-3 days requiring bilingual experts to under 10 minutes using basic computing resources), revolutionizing linguistic research efficiency through advanced natural language processing techniques. Published in Linguistic Society of America, a top linguistics conference.

## EDUCATION

- Georgia Institute of Technology — August 2024 – December 2025
  Graduate Researcher of Efficient Machine Learning Systems
- University of South Carolina — January 2024 – August 2024
  Degree: Accelerated Master's in Computer Science
- University of South Carolina — January 2021 – December 2023
  Degree: Bachelor's in Computer Science

## AWARDS / RECOGNITION

- SoftBank Group Representative for Entrance Ceremony 2026 — March 2026
- NSF GRFP Awardee — April 2024
- Toyo University Exchange Student Representative — July 2023
- University of South Carolina Outstanding Senior Class of 2023 — February 2023
- U.S. Department of State Critical Language Scholar of Japanese — January 2023

## QUALIFICATIONS

- Japanese Language Proficiency Test (N1)

## LANGUAGE SKILLS

- English – Native Level
- Japanese – Native Level (N1 Certified)
- Mandarin – Intermediate Level
- German – Beginner