

Lucas Cho

lucascho2000@gmail.com | +1 206-886-7549 | [linkedin.com/in/lucasthcho](https://www.linkedin.com/in/lucasthcho) | Portfolio

EDUCATION

Northeastern University (MS) — Major: Artificial Intelligence Jan 2026 – Dec 2027
Coursework: Machine Learning, Reinforcement Learning, Algorithms, Applied Programming for AI.

University of Washington (BS) — Major: Applied and Computational Math & Data Science Aug 2019 – Aug 2025
Coursework: Data Structures and Algorithms, Object-Oriented Programming, Database Systems, Software Engineering, Computer Systems, Distributed Systems, Testing & Quality Assurance.

WORK EXPERIENCE

Multimodal Machine Learning Engineer Intern | [Gobi](#) Aug 2025 – Present

- Fine-tuned Whisper with per-event LoRA adapters to adapt ASR to customer-specific jargon and named entities, building a failure-mining and contrastive-dataset pipeline that cut entity-level WER from 9.91% to 8.46% raw (5.58% with canonicalization), fixing 57 entity errors across five event domains.
- Improved speaker identification precision by 20% by architecting a matching system that fused audio diarization (ECAPA embeddings) with visual face-detection co-occurrence signals.
- Streamlined downstream LLM reasoning by engineering an end-to-end activity generation pipeline that converted raw transcriptions into structured format with associated confidence and ambiguity metadata.
- Reduced inference latency by 15% and increased throughput by designing an optimized preprocessing pipeline (VAD trimming, segment merging) and containerizing workers with environment-specific dependency groups.

Software Engineer | [Tiptop Technologies](#) Dec 2024 – Jul 2025

- Developed scalable backend APIs using Python/Flask supporting high-volume payment processing systems, improving API response time by 25% through code optimization and efficient data handling.
- Built robust data extraction pipelines using Python to process transaction data from multiple platforms, achieving 95% success rate in data collection while ensuring security compliance for financial systems.
- Collaborated with cross-functional teams to optimize system performance, reducing page load times by 15% through efficient debugging and troubleshooting of software applications in Agile development cycles.

Software Engineer Intern | [Dark Tonic](#) | [Blunderbusted](#) May 2024 – Dec 2024

- Led cross-functional team to design modular system architecture in C#, implementing best practices including unit testing and peer reviews, reducing integration time by 15% across builds.
- Improved system performance and memory optimization by 17% by implementing compression algorithms and version control solutions, demonstrating problem-solving and root cause analysis capabilities.

PROJECT / LEADERSHIP

Haven — Offline First AI-d Kit | [Gemma 4 “AI for Good” Hackathon \(Kaggle\)](#) May 2026

- Built an offline-first, on-device first-aid and safety assistant powered by Gemma that guides people through emergencies during war or disaster with no network connectivity.

High-Performance Computing Researcher | [Research Computing Club](#) Mar 2025 – Jun 2025

- Optimized multi-GPU training for deep learning models using PyTorch and TensorFlow, achieving 20% faster training time for large transformer models across multi-GPU clusters.
- Integrated AWS Neuron SDK with Inferentia and Trainium accelerators to optimize model inference speed, resulting in a 15% increase in inference efficiency on AWS hardware.
- Managed distributed workloads and parallel computing using Slurm and MPI on the Hyak supercomputing cluster, optimizing the training process for large-scale ML models.

NLP Research Analyst | [Seattle Children's Research Institute](#) Oct 2023 – Jun 2024

- Analyzed 900+ unstructured interview transcripts by applying k-means clustering and Latent Dirichlet Allocation (LDA) to identify latent patterns in homelessness and mental health outcomes, designing the full ML pipeline from data preprocessing to insight extraction.
- Engineered a sentiment analysis pipeline using Python (NLTK, TextBlob) to quantify emotional responses across mental health surveys, enabling data-driven identification of emotional triggers.
- Presented ML methodology and findings at the University of Washington Research Symposium, communicating complex NLP models and statistical results to technical and non-technical stakeholders.

SKILLS SUMMARY

Programming Languages: Java, Python, C++, C#, C, Go, JavaScript, TypeScript

AI: Codex, Gemini CLI, Lovable, Claude, Cursor

Database, Cloud & Web: SQL, AWS, Azure, CI/CD, Terraform, distributed storage systems, concurrency, React, Node.js