

Education

Johns Hopkins University

B.S., Computer Science

Baltimore MD

Expected 5/2026

- Coursework: Data Structures, Algorithms, Intermediate Programming (C, C++), Computer System Fundamentals (Assembly), Full-Stack JavaScript, Machine Translation (PyTorch), Databases (SQL)

Employment

Software Engineer, Intern

Microsoft

Redmond, WA

Feb 2024 – Present

- Accelerated LLM inference latency by 94.6% (from 57.1s to 3.1s) by integrating a custom TensorRT plugin, removing a critical performance bottleneck for multiple GPU-powered applications in Azure AI Services.
- Improved model accuracy by adjusting computation parameters in log bucket calculations, reducing output discrepancies between the original model and the plugin-enabled model by 98%, supporting more reliable inference results in production environments
- Designed and deployed a CI/CD pipeline in Azure DevOps to automate dependency management, which reduced manual engineering overhead by over 50 hours each quarter and mitigated security risks from outdated components.
- Engineered a novel multi-LoRA deployment strategy for quantized models, overcoming hardware operator limitations by implementing a zero-padding technique to simulate dynamic adapter ranking. This solution eliminated a critical model loading bottleneck and significantly improved service performance.

Co-Founder, Full Stack Developer

SummerNest

Baltimore, MD

Jul 2024 – Dec 2024

- Developed a React.js-based website for facilitating medium-term housing arrangements near college campuses, implementing features such as date filtering and location-based search
- Architected a Node.js and Express.js backend integrated with PostgreSQL
- Interviewed 20+ property owners and students to identify key pain points, leading to feature enhancements
- Accepted into Johns Hopkins' "Spark" start-up accelerator, receiving \$500 in grants for project development

Health Equity Data Scientist, Intern

Value Systems and Sciences Lab, UW Medicine

Seattle, WA

Jul 2022 – Sep 2023

- Leveraged R to analyze extensive data samples from a national physician survey
- Identified the effects of financial incentives on access to care for over 400,000 clinicians
- Findings presented at a national research meeting and published in a peer-reviewed journal

Projects

Attention-Based Neural Translator

- Optimized an NMT model processing speed by over 30% through batching experiments
- Enhanced translation accuracy by implementing beam search optimizations, achieving a 60% increase in BLEU score on test sets

Goalshare

- Developed an iOS app in Swift focused on goal setting and progress tracking
- Leveraged Firebase for real-time synchronization, robust authentication, and scalable cloud storage

Skills

- Languages: Java; C++; C; TypeScript; Assembly; SQL; JavaScript; Python; R; Swift;
- Technologies: Azure DevOps; ONNX Runtime; PyTorch; TensorRT; Git
- Concepts: Data Structures, Algorithms, Object-Oriented Programming, Large Language Models