

EXPERIENCE

Amazon | October 2024 - Present

Software Development Engineer 2, Personalization

- **Technical lead** for a cross-org **SLT goal** impacting **26M** customers: shipped a profile-to-account migration with shared Prime benefits, rearchitecting a **Tier-1** service to carry 100% of sign-out traffic, add **2,000 TPS**, and reach **99.99% availability**.
- Designed and shipped a production contextual-retrieval (**RAG**) API — an orchestration pipeline serving 100% of **Rufus** (Alexa-for-Shopping) traffic at **~50K TPS** and **~30ms p99**, using runtime query classification/enrichment and **in-memory KNN** to surface customer context to Gen-AI systems.
- **Technical lead** for profile support in Rufus (Alexa-for-Shopping) — built the profile-selection experience and enrolled **74M** Alexa profiles into Shopping via hybrid online/offline enrollment at **100%** success and no latency regression (**+19ms** hot path, under budget).
- Built **offline evaluation** and **golden datasets** that enabled a **fine-tuned decoder-only embedding model**, then led its client-invisible migration, improving **Average Precision +16 pts** and **F1@10 +6 pts**.
- Owned an **SLT goal** to distribute 8 EU marketplaces across a second AWS region for a **~15K TPS Tier-1** stack, improving resiliency and cutting the origin region's footprint **~60%** — **zero downtime** via gradual traffic shift.
- Replaced **manual, Sev2-triggered** ML-image patching with an automated pipeline (**CodePipeline** + a Python Lambda/CodeBuild) validating each image on **ephemeral SageMaker canaries** for model regressions before staged rollout — eliminating 3–4 security Sev2s/week.
- Designed a **Tier-1** identity API, cutting **p99** cold-start **~3,000ms** → **~7ms** at **1,000+ TPS** with a novel architecture to keep auth tokens warm under low traffic.

Amazon | June-September 2023

Software Development Engineering Intern, Personalization

- Built a global **purchase-recommendation system** for EU cross-marketplace shoppers; **multi-threaded** data access at **8ms p99**, validated by **A/B testing** that lifted paid units.

Simplified.io | June-September 2022

Software Engineering Intern

- Ported the **ezPulse** mobile app to web with **Flutter**, adding goal-creation and shipping a fully deployable site beyond the original POC scope.

Bryt Learning | July-September 2021

Software Engineering Intern

- Built a cross-platform **Flutter** classroom app and a **Django** admin tool (jQuery/AJAX) for timetable creation to onboard low-income schools.

PROJECTS

Semantic Compiler Frontend for LLM-Assisted Refactoring

Personal Project | 2024

- Built an **ANTLR**-based compiler frontend to transform large **Java** codebases into **semantic context trees**, enabling deterministic, **LLM-assisted** code edits for large-scale **technical debt reduction**, and shortlisted for a **YCombinator** interview

Semantic Code Graph for Observability Automation

Personal Project | 2024

- Designed and implemented a **Rust** library to construct a **semantic code graph** using **Tree-Sitter** to embed large, language-agnostic repositories.
- Enabled **deterministic observability improvements**, including **automated OpenTelemetry instrumentation** and **structured logging** across large codebases.

Transactional Key-Value Store w/ Paxos

Class Project | 2023

- Designed and implemented a **sharded transactional key-value store** with **dynamic load balancing** and **multi-key atomic updates** across server groups.
- Used the **Paxos** consensus algorithm to ensure **fault tolerance** and maintain progress despite node and network failures.

Multithreaded Search Engine & Web Server

Class Project | 2021

- Developed a **multithreaded web server** in C/C++, hosting a **search engine** with a **query processor** and **directory crawler**, and built an **on-disk inverted index** to support efficient document indexing and retrieval.

Interpreter for a Python-like Language

Personal Project | 2020

- Designed a weakly-typed programming language supporting **control flow**, **functions**, and **recursion**, and implemented a Python **interpreter** including **lexing**, **parsing**, **AST construction**, and **execution**.

EDUCATION

University of Washington - Bachelor of Science, Computer Science (CS) - 3.92/4.00 (Cum Laude)

June 2024

SKILLS

- **Languages:** Java, TypeScript/Javascript, Python, Rust, C/C++ | **Frameworks:** gRPC, Spring Boot, Node.js, React, React Native, OpenTelemetry, Tokio | **Engineering & Productivity:** Git, CI/CD, Docker, k8s, AWS, IaC