BINIYAM GEBREYOHANNES

biniyamgebreyohannes@gmail.com | LinkedIn.com/in/Biniyam | biniyamgebreyohannes.com | Seattle, WA

EDUCATION

University of Washington - Seattle **Expected Graduation Date: December 2026**

B.S. in Computer Science

- **Core Coursework:** Data Structures & Algorithm Design, **Object Oriented Programming**, Operating Systems(memory/resource management), **Distributed Systems, Database Systems**, Linear Algebra, **Query Processing, Scheduling and Process Control**

TECHNICAL SKILLS & SOFTWARE PROFICIENCIES

Languages: Python, C/C++, Rust, Java, Kotlin, SQL, TypeScript, JavaScript

Systems & Infra: Linux, Bash, **Docker, Kubernetes, SLURM**, AWS, Azure, S3, Datadog, Git, GitLab

Frameworks & Platforms : Node.js, React, Next.js, RESTful APIs, Microservices

ML : **PyTorch**, **CUDA**, Pandas, OpenCV, YOLOv6, scikit-learn

EXPERIENCE

***Incoming SWE Intern*, Adobe** **May 2026 - August 2026**

- Incoming Software Engineering Intern at Adobe, joining the team building **LLM gateway** infrastructure for production **AI workflows**, with focus on routing, usage controls and scalable **Kubernetes** deployment.

***SWE Intern*, Expedia** **June 2025 - August 2025**

- Built a replay-driven traffic load tester and validation framework for Spring Cloud Gateway and Envoy, enabling early detection of performance regressions in platform handling **60M+** monthly API transactions and sub-20ms latency at **99.99%** uptime.
- Built observability and validation tooling with **Kotlin, Datadog,** and **OpenTelemetry** to trace request flows, surface outlier behaviors, and automate deployment checks, reducing triage time by 40% and speeding incident recovery during gateway releases.

***Undergraduate Researcher*, User Empowerment Lab** **January 2025 - August 2025**

- Published **AAAI 2025** paper, Understanding Privacy Norms Around LLM-Based Chatbots: built privacy-aware chatbot interface with input masking and consent handling, deployed to **300+** users .
- Engineered and analyzed **9,000+** contextual integrity vignette interactions through the chatbot system to study user privacy expectations across conversational contexts.

***AI/ML Engineering Intern*, STEMTAC** **January 2024 – June 2024**

- Led a **20-membe**r team across two phases, each lasting two months, to redesign a real-time facial recognition pipeline using **PyTorch**.
- Redesigned a real-time face detection pipeline using **OpenCV** and **YOLOv6**, improving accuracy by 25% on **5,000+** frames.

PROJECTS

Dynamic Batching Inference Engine | Project **March 2025**

- Built a **Python based** inference service with a centralized request queue and **asynchronous batching pipeline to** handle high-concurrency workloads and improve throughput.
- Designed and evaluated scheduling strategies (FIFO vs. adaptive batching), measuring **latency–throughput tradeoffs**, queue wait times, and resource utilization across local and **SLURM**-scheduled experiment runs.
- Benchmarked **CPU vs GPU** execution and analyzed memory usage and batch size scaling to achieve 2–3× throughput gains while maintaining low p95 latency in a containerized deployment.

Rust Traffic Replayer (CLI) | Project **October 2025**

- Built a high-throughput **Rust CLI** that replays **NDJSON HTTP** logs with Tokio and reqwest, streaming from **S3** and applying PII sanitization plus method/path/status filters via clap.
- Sustained **~2–3k req/s** to a public endpoint **(p99 150–200 ms)**, **~5k req/s** to a local mock service, and **~150 GB/day** from **S3**.

Dr. AI Assistant | Project **September 2024**

- Built and deployed a **Python + Flask** backend serving a fine-tuned **LLaMA** model, enabling 5,000+ remote consultations and reducing patient-to-doctor ratio by **89%** in under-resourced clinics.
- Designed a shared-nothing architecture for horizontally scalable inference across CPU-only nodes, enabling deployment in **GPU**-constrained environments.
- Improved diagnostic accuracy by **25%** via LLM-driven feedback loops and structured prompt refinement.

LEADERSHIP & ACTIVITIES

Hackathon Competition, DubHacks '24 **October 2024**

- Won Honorable Mention at **DubHacks** 2024 for building a GPT-powered educational assistant to answer student questions .