Ansel K. Erol

linkedin.com/in/ansel-erol/ | github.com/A-K-Erol

Education

Georgia Institute of Technology | *B.S. & M.S. in Computer Science* ⋄ GPA 4.0 — *May 2026*
Graduate Teaching Assistant: Introduction to Artificial Intelligence — Atlanta, GA

Professional Experience

Baseten Labs — *Starting May 2026*
Incoming Forward Deployed Software Engineer — San Francisco, CA

Google — *May 2025 – Aug. 2025*
Software Engineer Intern — Mountain View, CA

- Optimized a low-latency C++ inference system that delivers relevant ads without sacrificing user privacy.
- Enhanced scalability by transitioning to an async request model, boosting model serving throughput by 1.6x.
- Slashed deployment costs by $800k/yr by implementing cancellation propagation and deadline enforcement in gRPC services, enabling usage of 75% smaller virtual machines while maintaining inference performance.
- Developed an automated performance testing pipeline on GCP, leveraging Terraform Infrastructure-as-Code and synthetic load with wrk2 to validate system reliability and identify bottlenecks before deployment.

Georgia Tech ML Infrastructure and Architecture Lab — *Jan. 2024 – Present*
Research Assistant — Atlanta, GA

- Formulated a real-time simulation protocol to assess model training and inference bottlenecks, leveraging multithreaded C++ to simulate parallel communication/computation and Python to analyze data.
- Architected a system for prioritizing satellite images for downlink utilizing an EfficientNet backbone and task-specific prediction heads, pipelining and distributing load between the CPU and an AI accelerator.
- Engineered a dynamic, power-aware scheduler for prioritization tasks that reduces latency by 2.1x through eliminating redundant inferences, which I presented as first author at MLSys ’25 and ’26.
- Modeled the joint bottlenecks of expert distribution across GPUs in MoE Language Models (i.e., DeepSeek, GPT-OSS) as a Mixed-integer Program and solved it via Tabu Search, yielding 8-10% higher throughput.

Travelers Insurance — *June 2024 – Aug. 2024*
Software Engineer Intern

- Automated the validation of 1300 tables migrated to AWS S3, reducing validation time by 80%.
- Reduced data volume in ~300m record table by 65% through a retention policy, enhancing data governance.
- Created a new dashboard for tracking policy renewals, now in production use by over 700 account executives.

People for PSEO — *Sept. 2022 – Aug. 2025*
Finance Director & Board Member — Minneapolis, MN

- Oversaw finance committee and led successful grant and donor initiatives, resulting in 2.3x budget growth.
- Formed and maintained partnerships with non-profits, student organizations, and the MN Dept. of Education.

Projects

- **ShopEase:** Microservices-based e-commerce platform with Spring (Java) backend that provides secure product management and order tracking via OAuth 2.0, deployed on AWS using Lambda and RDS (Postgres).
- **Workforce Scheduling and Routing:** Gradient-boosting classifier that predicts the best optimization algorithm given a problem instance and runtime constraint, achieving 40% higher performance than any algorithm individually; resulting co-author paper is in review.
- **MeloDEX:** AI-powered search engine that lets users find songs by simply describing the mood or style, incorporating a sleek UI, relational filtering, efficient vector search, and Spotify and Apple Music integrations.

Skills

Programming: Python, C++, Java, C, SQL, JavaScript, Bash
Technologies: Git, GitHub, Docker, Linux, SQL Server, Excel, Postgres, REST APIs, Azure, AWS, Google Cloud
Machine Learning: PyTorch, TensorFlow, SciKit, SciPy, Pandas, NumPy
Interests: Backend Design, Distributed Systems, High Performance Computing, Edge Computing