Anand Tandon

+1 647-904-2370 | anand.tandon@uwaterloo.ca | linkedin.com/in/anandtandon8 | github.com/anandtandon8

Education

University of Waterloo Expected Grad. Apr 2028

Bachelor of Software Engineering (BSE) *Waterloo, ON*

- Courses: Algorithms, Databases, Compilers, Computer Architecture, Object Oriented Programming

Skills

Languages: Python, JavaScript/TypeScript, Go, C, C++, SQL, Bash, Java

Tools: AWS, Google Cloud, Docker, Kubernetes, Next.js, React, Flask, PostgreSQL, Linux, Git, nginx

Experience

Ideogram Jan 2026 - Apr 2026

Software Engineering Intern *Toronto, ON*

- Recovered ∼**$3.2M/year** in idle compute time by building a Flask/React + Spanner DB job scheduler on top of the limited **GCP** TPU API to manage ∼300 concurrent training and inference jobs
- Sped up fine-tune loads **5x** and inference latency **1.1x** by saving fused LoRA in HBM for weight resets via JAX
- Diagnosed and **resolved critical GPU hangs** in inference workers through inter-GPU communication analysis (NCCL debugging, py-spy, JAX) and collaboration with the Digital Ocean and AMD teams
- Cut inference costs by ∼**$3M/year** by leading migration from GCP TPUs to **AMD GPUs** on Digital Ocean K8s, using Megaport for multi-cloud and a custom Prometheus/K8s based dashboard for observability
- Built a virtual clothing try-on API for a top enterprise customer using **Gemini** VLM + **SAM3** for segmentation

Toyota May 2025 – Aug 2025

Software Engineering Intern *Toronto, ON*

- Reduced hiring time by ∼**40%** by building a fullstack resume parser to automate resume screening
- Enhanced resume scoring accuracy by **70%** with semantic skill-matching using word embeddings and kNN
- Cut deploy time by **70%** by implementing Azure CI/CD with end-to-end testing on staging infrastructure
- Reduced monthly server maintenance workload from **30 hours to 2 hours** by designing and implementing an AWS auto-patching solution for OS and software package updates across **100+ EC2 instances**

Meta May 2025 – Aug 2025

MLH Production Engineering Fellow *New York City, NY (Remote)*

- Slashed production error rates by **25%** by improving **CI/CD** tests and implementing agentic review workflows
- Cut incident response time by **50%** for blog site by improving observability with Prometheus + **Alertmanager**

Projects

ATPhoto | **Portfolio** *Next.js, AWS, Docker, Firebase, PostgreSQL, nginx, TS* |

- Engineered an image processing pipeline from Adobe Lightroom exports to the portfolio site that automatically sorts and deploys images to categorized galleries using a custom Adobe integration, **PostgreSQL**, and **EC2**
- Commissioned **$2,500+** of professional photography shoots through visibility gained from the portfolio site
- Fine-tuned a ResNet-18 image classifier with ∼**97%** accuracy with **PyTorch** and deployed it to AWS Lambda

Cook-Buddy *Llama 3.1, OpenAI, Flask, Python, Google Cloud, JS* |

- Developed an AI-powered cooking assistant with **Llama 3.1**, using Unsloth with LoRA to achieve **2x faster fine-tuning** and reduce VRAM usage by **70%** during training compared to Hugging Face + Flash Attention
- Configured a Raspberry Pi to take sensor data and make REST API calls to a Flask server running the LLM