

Anshuman Sahoo

FOUNDATION MODELS · AGENTIC SYSTEMS · SCIENTIFIC AI

(416) 838-5337
anshuman264@gmail.com
linkedin.com/in/anshu92
Toronto, ON

PROFILE

Principal ML engineer with 9+ years taking scientific AI and foundation-model research into production. Delivered up to 16× higher training efficiency on targeted workloads, scaled a reusable 32B SFT pipeline across tens of GPUs, and led agentic and reinforcement-learning R&D for scientific reasoning.

CORE FOCUS

Training & Alignment
Megatron-LM · DeepSpeed · SFT · GRPO
Systems & Inference
Ray Data · vLLM · Docker · Kubernetes
Agents & Evaluation
KG-RAG · LLM-as-a-judge · VLM

EXPERIENCE

AUTODESK · Principal ML Engineer, Foundation Models (AEC)

NOV 2025 — PRESENT

- Rebuilt Ray-based pretraining infrastructure for per-batch data blending, replay, and VLM workloads; achieved up to **16× higher training efficiency** on targeted workloads through sequence packing, custom performance callbacks, and time-to-first-step optimization.
- Containerized Megatron-LM and published it to an internal registry, preserving researchers' existing training workflows while deploying a reusable SFT pipeline for a **32B-parameter model across tens of GPUs**; integrated ms-swift as a shared backend and recipe layer.
- Designed a knowledge-graph-grounded, multi-task data pipeline from raw AEC documents, improving foundation-model semantic and spatial reasoning over complex 2D documentation and floor plans.

BENCHSCI · Lead R&D Machine Learning Engineer

SEP 2024 — OCT 2025

- As **ML R&D lead**, partnered with scientific and data executives to de-risk early LLM and agent-platform architecture decisions; built **knowledge-graph RAG** that traversed disease-biology paths to generate novel hypotheses and developed LangGraph agents for experiment recommendation, literature triage, and scientific Q&A, with RLHF/GRPO post-training and LLM-as-a-judge evaluation.
- Pretrained and contrastively fine-tuned a scientific-retrieval embedding model to **MAP 0.52** on BenchSci's internal benchmark, outperforming three top-ranked MTEB models; deployed low-latency, high-throughput inference with vLLM and Ray.

Senior ML Engineer II · Discovery Tech Lead

AUG 2022 — SEP 2024

- Invented a syntax parser that converted unstructured text into knowledge-graph entities, increasing graph connectivity by **32%** and helping secure at least **two enterprise contracts**.
- Built and scaled a distributed multimodal LLM pipeline for scientific-figure extraction using data distillation; developed Western Blot quantification validated by correlation analysis, adding **30k+ graph connections**.

ML Engineer → Senior ML Engineer

FEB 2017 — AUG 2022

- As BenchSci's first ML engineer, shaped ML strategy; built GCP Dataflow/Vertex AI pipelines and managed **Google Cloud Container Registry** workflows for Apache Beam jobs; developed LSTM/CNN document systems and patented scientific figure-parsing technology (CA 3081623).

EDUCATION

B.A.Sc. in Computer Engineering

University of Toronto · 2016 | Bioengineering minor · Entrepreneurship certificate

EARLIER ENGINEERING

Neutun Labs · Analog Devices · Bhabha Atomic Research Centre

Seizure-detection R&D · digital design · research, 2013–2016