

AMAN SINGH

ML Systems Engineer | LLM Training, Serving & AI Infrastructure

2amansingh2@gmail.com | github.com/amanyagami | linkedin.com/in/amanyagami

EXPERIENCE

Noah Labs — Founding Engineer

May 2026–Present

Palo Alto, CA

- Build and operate **Sentinel**'s LLM/agent execution path for local and air-gapped environments, implementing model integration, repository context, multi-step tool orchestration, backend services, and evaluation.
- Implemented Sentinel's async agent runtime with streaming tool-calling loops, semaphore-bounded concurrent task execution, approval gates, and crash recovery for backend services.
- Hardened local execution sandboxing in **Rust** and upstreamed **arapuca** fixes covering Windows AppContainer/LPAC confinement and Linux filesystem isolation.

Amazon — Software Development Engineer

Jun 2022–Mar 2023

Bangalore, India

- Owned peak-planning and reliability systems for high-volume infrastructure, supporting a **1.8× increase in client traffic** while reducing operating costs by **30%**.
- Improved CI/CD deployment processes, reducing production incidents by **60%**.

Make Programming Simple Lab, Arizona State University — Graduate Researcher

Mar 2024–May 2026

Tempe, AZ

- Developed **Viyog**, a PyTorch framework for separating adversarial and out-of-distribution inputs using intermediate-representation geometry.
- Trained and evaluated models on **NVIDIA H100 GPUs** using PyTorch DDP and FP16/BF16 mixed precision with reproducible benchmark pipelines.
- Optimized GPU inference to **sub-10 ms latency on NVIDIA A6000**, profiling model behavior across robustness and OOD workloads; collaborated with Yonsei University on two papers under review for ESWEK 2026.

RESEARCH

Viyog: Separating Adversarial and Out-of-Distribution — Submitted to ESWEK CODES 2026; under review

OPEN SOURCE — ML SYSTEMS & INFRASTRUCTURE

SGLang · VERL · Megatron-LM · PyTorch AO · LLVM/MLIR · MLX-LM · arapuca

- Authored contributions/fixes spanning SGLang's **KV/radix-cache correctness**, VERL's **FSDP checkpoint integrity** and distributed vLLM rollout networking, and Megatron-LM pipeline-parallel configuration validation.
- Contributed quantization/FP8 and tensor/runtime correctness fixes in PyTorch AO; delivered **2 merged LLVM/MLIR fixes** for production compiler/runtime failures.

TECHNICAL SKILLS

Languages: Python, C++, Rust, Go

ML Systems: PyTorch, CUDA, DDP/FSDP, Megatron-LM, FP16/BF16, distributed training, quantization, evaluation

LLM / Inference: SGLang, vLLM, Transformers, RAG, agentic systems, LoRA/QLoRA, GPU inference

Infrastructure: Linux, Kubernetes, Docker, AWS, Google Cloud Platform, CI/CD, Spark

EDUCATION

Arizona State University — M.S. Computer Engineering

Jan 2024–May 2026

Thesis: *Viyog: Separating Adversarial and Out-of-Distribution*

National Institute of Technology Karnataka — B.Tech Electrical and Electronics Engineering

2018–2022