# AMAN SINGH

Palo Alto, CA

Head of Engineering | ML Systems, LLM Agents & Applied Research

github.com/amanyagami | linkedin.com/in/amanyagami | aman-ai-portfolio.netlify.app

Hands-on engineering leader building Sentinel and ARK, with Amazon production-systems experience and ASU research spanning reliable ML, GPU inference, and efficient model adaptation.

## EXPERIENCE

**Noah Labs | Palo Alto, CA** May 2026 - Present

**Head of Engineering** Aug 2026 - Present

Founding Engineer May 2026 - Aug 2026

- Lead engineering and hands-on development of **Sentinel and ARK**.
- Build Sentinel's LLM/agent execution path for local and air-gapped environments, spanning model integration, repository context, tool orchestration, backend services, and evaluation.
- Implemented an async runtime with streaming tool calls, bounded concurrency, approval gates, and crash recovery.
- Hardened Rust sandboxing; upstreamed Windows AppContainer/LPAC and Linux filesystem-isolation fixes to arapuca.

**Arizona State University | MPS Lab** Mar 2024 - May 2026

Graduate Researcher | Tempe, AZ

- Developed **Viyog**, a PyTorch framework separating adversarial and out-of-distribution inputs using representation-level signals; collaborated with Yonsei University.
- Optimized A6000 inference to **sub-10 ms latency**; trained and evaluated models on H100 GPUs with PyTorch DDP and FP16/BF16 mixed precision.

**Amazon | Bangalore, India** Jun 2022 - Mar 2023

Software Development Engineer

- Owned peak-planning and reliability systems, supporting **1.8x client traffic** while reducing operating costs by **30%**.
- Improved CI/CD deployment processes, reducing production incidents by **60%**.

## SELECTED RESEARCH

**Viyog: Separating Adversarial and Out-of-Distribution** | ESWEEK / CODES 2026; first author, accepted.

**EKLAV - Reranking Distillation** | Submitted to ICLR 2027.

Teacher-context, answer-only distillation on BRIGHT; manuscript reports up to **31.6% fewer estimated training FLOPs** and **32.7% less inference time** than full-trace distillation.

## OPEN-SOURCE CONTRIBUTIONS

- **SGLang, VERL, Megatron-LM:** KV/radix-cache correctness, FSDP checkpoint integrity, distributed vLLM rollout networking, and pipeline-parallel configuration validation.
- **PyTorch AO, LLVM/MLIR:** quantization, FP8, and runtime-correctness fixes; **2 merged LLVM/MLIR fixes** for compiler/runtime failures.

## TECHNICAL SKILLS

**Languages:** Python, C++, Rust, Go.

**ML systems:** PyTorch, CUDA, DDP/FSDP, FP16/BF16, LoRA/QLoRA, quantization, evaluation.

**Serving & infrastructure:** SGLang, vLLM, Transformers, RAG, Linux, Docker, Kubernetes, AWS, GCP, CI/CD.

## EDUCATION

Arizona State University | M.S., Computer Engineering Jan 2024 - May 2026

NIT Karnataka | Bachelor's, Electrical and Electronics Engineering 2018 - 2022