Shiven Saini

Portfolio: shiven.dev

+91-740-427-5751 shiven.career@proton.me LinkedIn GitHub

Technical Skills

Programming Languages: Python, C/C++, Kotlin, TypeScript
ML/AI Frameworks: PyTorch, TorchRL, Stable-Baselines3, PEFT, LoRA, QLoRA, YOLOv11, Langchain, OpenCV
Robotics & Simulation: PyBullet, MuJoCo, Gazebo, ROS2, URDF, COLMAP, GLOMAP, Gaussian Splatting
Technologies & Frameworks: CUDA, Next.js, Qt6, Spring Boot, Ktor, ollama, vLLM
Tools & Platforms: Git, AWS (EC2, SageMaker), Docker, Linux, Android Studio, Visual Studio
Specializations: AI/ML, Robotics Simulation, GPU Programming, Full-Stack Development, Android Development

Experience

Backend Developer Dec. 2025 – Present
Stringly.AI *Remote*

- Architected AI-powered recommendation system for intelligent date profile matching based on user intentions, implementing custom scoring algorithms to quantify compatibility metrics and match quality
- Developed hybrid auto-reply suggestions system leveraging RAG (Retrieval-Augmented Generation) architecture and Langchain framework for contextually-aware conversational responses
- Optimized existing AI workflow infrastructure achieving 18-20% reduction in API calls and GPU compute costs through efficient request batching, caching strategies, and CUDA toolkit deployment optimization
- Fine-tuned and deployed custom PEFT LoRA models based on Nemotron3-nano for domain-specific language tasks, significantly improving response quality while maintaining low latency

Robotics Simulation & ML Engineer Contract Sept. 2025 – Dec. 2025
Shenzhen Robora Automatic Robots Ltd Enterprise — GitHub *Remote*

- Architected comprehensive VLA (Vision-Language-Action) SDK with multi-model inference and fine-tuning support for SmolVLA, Pi0, Pi0.5, and Groot N1.5 using PEFT techniques (LoRA, QLoRA) for both action head and full model optimization
- Designed and implemented multi-physics simulation environments using PyBullet, MuJoCo, and Gazebo for robot model validation, integrating classical control (PID, LQR, MPC) and modern RL-based controllers via Stable-Baselines3 and TorchRL
- Developed 3D reconstruction toolkit leveraging photogrammetry techniques (Structure from Motion, COLMAP, GLOMAP) and 3D Gaussian Splatting for ultra-realistic environment modeling and scene reconstruction
- Engineered GDPR-compliant data anonymization pipeline by fine-tuning custom YOLOv11-medium model achieving 95%+ accuracy in detecting and blurring faces and license plates for EU privacy compliance

AI/ML Software Engineer Intern July 2025 – Sept. 2025
Uniconverge Technologies Pvt Ltd. *Noida, India*

- Developed native Windows application using C++ and Qt 6.5.3 for comprehensive LAN IP camera system management with RTSP streaming capabilities
- Integrated AWS SageMaker AI for real-time employee productivity monitoring through intelligent camera stream analysis
- Implemented Wireguard VPN tunneling with TCP port forwarding for secure peer-to-peer communication with AWS EC2 infrastructure
- Designed system-level features including system tray integration and background service functionality

Key Projects

PithuuOS - Performance Linux Distribution — GitHub — *Linux Kernel, C/C++, Python* 2020 – Present

- Built performance-optimized Linux distribution tailored for developers with pre-configured development suite and utilities
- Designed custom package manager infrastructure using PKGBUILD scripting with automatic builds and dependency resolution
- Provided separate builds for AMD/Intel and NVIDIA architectures; enhanced performance through kernel parameter optimization

cuRay-Tracer - CUDA ray tracing engine — GitHub — *C++17, CUDA, OpenGL* July 2025

- Built a real-time CUDA-accelerated ray tracer with 7-level depth support for interactive 3D rendering on NVIDIA GPUs utilizing OpenGL for display.
- Implemented interactive camera controls with mouse navigation and dynamic light source control via keyboard inputs.
- Supported cross-platform deployment on Linux, Windows, and macOS with high-performance GPU ray tracing kernels.
- Designed modular code architecture separating rendering kernels, camera management, and scene setup for clarity and extensibility.
- Provided detailed troubleshooting guidance and performance tips to support easy deployment and debugging.

Visco Connect - Windows Native RTSP Forwarding Software — GitHub — *C++17, Qt 6, CMake* July 2025

- Architected and implemented a modular native Windows LAN camera port-forwarding application that enables seamless RTSP stream redirection from multiple network cameras to local ports, supporting auto-assigned dynamic port mapping.
- Designed core components including CameraConfig (metadata management), CameraManager (stream orchestration), and PortForwarder (bidirectional TCP tunneling), with clean separation of responsibilities for scalability and testability.
- Leveraged Qt networking primitives, TCP sockets, and JSON-based APIs for robust configuration management and runtime service control, ensuring safe cross-platform operation with RAII-based resource safety.
- Developed a system tray utility and QtWidgets GUI for real-time monitoring, logging, and interaction, including background Windows service deployment with headless start and diagnostics.
- Established reproducible builds with CMake, rigorous logging through a custom Logger, and implemented comprehensive error handling and low-latency, reliable network performance.

Education

Deenbandhu Chhotu Ram University of Science and Technology Sonipat, Haryana

Bachelor of Technology in Electronics & Communication Engineering — IoT Specialization *Nov 2022 – Present*

Achievements & Recognition

E-Yantra Robotics Competition — *Team Leader* 2023, 2024

- **2024:** Led autonomous warehouse drone development using ROS2, OpenCV, and Gazebo simulation
- **2023:** Developed self-balancing lunar scout robot with PID/LQR control systems and Fusion 360 design

Smart India Hackathon — *Team Leader* 2023, 2024

- **2024:** Water conservation cross-platform game using Godot Engine — **2023:** Ubuntu Linux hardening solutions

Certifications

Programming & Development:

- Meta: Advanced Programming in Kotlin
- John Hopkins: GPU Programming Specialization
- Linux Foundation: Open Source Software Development, Linux and Git
- NPTEL: Programming in Modern C++

AI & Cybersecurity:

- IBM: RAG and Agentic AI Professional Certificate
- Google: AI Essentials Professional Certificate
- Duke University: Large Language Model Operations
- Google: Cybersecurity Professional Certificate
- NVIDIA: AI Infrastructure and Operations Fundamentals