RAJ SINGH CHAUHAN

Software Engineer | Backend, Distributed Infrastructure & ML Systems
rajsinghchauhan352@gmail.com | linkedin.com/in/rajchauhan28 | github.com/rajchauhan28 | rajchauhan28.site

EDUCATION

Amity University Rajasthan — Jaipur, India
Bachelor of Technology in Computer Science; CGPA: 7.69 — 2022 – 2026

TECHNICAL SKILLS

Languages & Backend: Python, Rust, C++, Bash, SQL, Django, PostgreSQL, RESTful APIs, API Development
Cloud & Infrastructure: Linux, Proxmox VE, Proxmox Backup Server, AWS, Oracle Cloud (OCI), Docker, CI/CD
Systems & Architecture: Distributed Systems, Virtual Desktop Infrastructure (VDI), HA Clustering, SDLC
ML & Inference: PyTorch, Hugging Face, Triton Kernels, Transformers, Quantization (INT8, 4-bit, 1.58-bit)

EXPERIENCE

ABS Technologies — India
Software Developer (Infrastructure & Virtualization Systems) — May 2026 – Present

– Deployed and managed a multi-site Virtual Desktop Infrastructure (VDI) on **Proxmox VE**, provisioning centralized Windows VMs for business users across geographically separated branch nodes.
– Configured Corosync cluster quorum, scheduled cross-node storage replication, and automated high-availability (HA) failover policies to keep distributed Windows VM workloads synchronized.
– Automated incremental VM snapshot schedules and off-site disaster recovery pipelines using **Proxmox Backup Server (PBS)** integrated with **AWS** cloud storage across the SDLC.

Bluestock Fintech — Remote
Backend Developer Intern (Led 4-member module team) — May 2025 – Jun 2025

– Led a 4-member backend team to build a high-concurrency investment platform using **Django**, **PostgreSQL**, and **RESTful APIs** with connection pooling and modular service layers.
– Eliminated database bottlenecks via composite indexing, relational query optimization, and response caching, cutting average API latency by ~20% under concurrent load testing.
– Integrated OAuth 2.0 authentication and role-based access control across backend endpoints to secure third-party API throughput.

Outlier.ai — Remote
AI/Systems Evaluator (Python & SQL Code Optimization) — Jul 2023 – Jan 2024

– Audited and benchmarked AI-generated **Python** and **SQL** code submissions for algorithmic time/space complexity, query execution efficiency, and boundary edge-case handling.
– Authored reference systems-programming solutions and structured RLHF evaluation rubrics to fine-tune foundational code models and reduce reasoning failure modes.

PUBLICATIONS

NanoBit: Extreme Compression for Vision Architectures via 1.58-Bit Ternary Quantisation — Oct 2026
IET Conference Proceedings, Vol. 2026, Iss. 15, pp. 791–798 (ICRACS 2026) — First Author — DOI: 10.1049/icp.2026.4002

– Formulated a hybrid-precision vision architecture with 1.58-bit ternary ({−1, 0, 1}) pointwise and FP32 spatial layers, achieving 82.83% top-1 accuracy on ImageNette-320 in 9.98 MB (10.7× smaller than 106.8 MB FP32).

KEY PROJECTS

NanoBit Edge Inference Runtime & Bit-Packing Kernels (GitHub) — *Rust, C++, Triton, PyTorch*

– Engineered 2-bit bit-packing kernels in **Rust** and custom **Triton** compute kernels with a **C++ inference engine**, cutting low-precision arithmetic and dequantization overhead by 70% vs. FP32 execution.
– Deployed the quantized runtime on **Raspberry Pi 5** (ARM Cortex-A76), achieving 6.7× faster inference (3.13 FPS) and >20× runtime RAM reduction (<100 MB peak memory vs. >2 GB FP32 baseline).

High-Throughput Quantized Inference Pipeline (GitHub) — *Python, PyTorch Internals, INT8 Quantization*

– Implemented dynamic **INT8 weight-only quantization** and custom memory hooks via **PyTorch internals**, cutting model memory by **40%** vs. FP16 baselines and sustaining **~50ms/token** latency on consumer hardware.

Hypr-Dotfiles: Automated Linux Provisioning (GitHub) — *Bash, Linux (Ubuntu/Debian), Automation*

– Built idempotent **Bash** provisioning scripts for Ubuntu/Debian environments, cutting machine setup time by 80% (from ~2.5 hours of manual configuration to <30 minutes via one-command bootstrap).

CERTIFICATIONS & ACHIEVEMENTS

Certifications: OCI Generative AI Professional (2025), Amazon ML Summer School (2024), HackerRank Python (5 Star)
Awards: 1st Place, WhackHackathon 2024 (Amity University); Finalist, IIT Delhi Techfest (Syntaize AI); 1st Place, Robo Race.