Rohit Yadav

Built 25+ LLMs from scratch • 120M–1.4B params • TPU clusters • Trained on 500 Billion+ tokens

+91 8849994303 | yrohit1825@gmail.com | linkedin.com/in/rohit-yadav-25535b256 | github.com/YADAV1825 | rohit-portfolio-yadav1825.vercel.app

Education

National Institute of Technology, Jalandhar — Jalandhar, India
Bachelor of Technology in Information Technology — *Aug 2023 – Aug 2027*

Green Valley School for Children — Gandhinagar, India
CBSE Board 12th — *May 2021 – May 2023*

Professional Summary

Third-year undergraduate and Systems + AI builder who enjoys building complex systems from scratch. Built 25+ LLMs from scratch (120M–1.4B parameters) on Google's TPU clusters, working with 10TB+ datasets (PILE, DOLMA) and 500B+ tokens. Experienced in **pretraining, fine-tuning, and continual pretraining** of large-scale models. Designed high-throughput data pipelines and tokenization workflows for distributed training. Experience developing clinical-scale genomic ML systems and performance-critical infrastructure. Awarded Google TPU Research Cloud (TRC) grant and AMD MI300X compute access for large-scale experimentation.

Experience

Founder & Research Lead — AutonomousX — Apr 2026 – Present
Open Source LLM Research Organization

- Compute supported via **Google TPU Research Cloud (TRC)** grant, enabling large-scale LLM training experiments.
- Trained **25+ LLMs (120M–1.4B params)** over **500B+ tokens** using TPU pods (v4-8 to v4-32, v6e-64, 1TB+ HBM/VRAM).
- Engineered highly optimized JAX training pipelines scalable from **small TPU pods (v4-8)** to multi-host distributed setups.
- Built reproducible training systems allowing end-to-end LLM development without reliance on large GPU clusters.

Open Source Contributor — 2026 – Present
Meta LLaMA and Google DeepMind Gemma

- Contributed to production-scale LLM repositories by fixing critical bugs and improving usability.
- **Google DeepMind Gemma** — resolved tokenizer normalization and jax typing issues.
- **Meta LLaMA** — fixed incorrect CLI import causing ModuleNotFoundError in download pipeline.
- Worked with large-scale transformer codebases involving **training, inference, and tokenization pipelines**.

Full-Stack Web Developer Intern — Jun 2025 – Jul 2025
Skylark Express Delhi Pvt Ltd — *On-site, Gurugram*

- Developed internal logistics dashboard improving data visibility and workflow efficiency.
- Built frontend using **React, Next.js, TypeScript** and integrated **REST APIs**.
- Improved UI responsiveness and system performance for internal operations.

Projects

AutonomousX: TPU-Native LLM Training Organization (Instinct Model Family) [HuggingFace]

- Built and open-sourced **25+ LLMs (120M–1.4B params)** by engineering **TPU-native distributed pipelines (JAX pmap)** under AutonomousX.
- Trained the **Instinct model family** on **500B+ tokens** (PILE, DOLMA) under extreme overtraining regimes to study **scaling laws, convergence, and degeneration dynamics**.
- Open-sourced complete pipelines including HuggingFace weights, training logs, and a 10-minute TPU setup guide enabling full reproducibility for researchers.
- Provided a ready-to-run inference notebook allowing quick evaluation and usage of the trained models.

PathoPreter: Clinical-Grade SNV Pathogenicity Ranker (DNA Foundation Model) [HuggingFace]

- Developed a **hybrid genomic foundation model (500M params)** fusing a Nucleotide Transformer backbone with clinical tabular features for **SNV pathogenicity prediction**.
- Achieved **0.9186 ROC-AUC** on a rigorous 14k rare-variant benchmark, outperforming industry SOTA (AlphaMissense, CADD) by **+0.32 ROC-AUC**.
- Designed a **ranking-first clinical triage system** demonstrating **75%+ pathogen recall** while testing only 5–10% of variants, drastically reducing lab turnaround time and costs.
- Validated biological learning via **SHAP ablation studies** (showing 64.9% reliance on raw DNA) and open-sourced the **end-to-end reproducible pipeline**.

RohitOS: Custom Operating System (C + x86 Assembly) [GitHub]

- Built a **minimal operating system from scratch** with custom bootloader and 32-bit protected mode kernel.
- Implemented low-level components including **GDT setup, disk loading (INT 13h), and kernel entry**.
- Developed a **CLI shell with command parser**, system utilities, and custom memory-safe string/IO libraries (no libc).
- Designed complete build system using **NASM, GCC (-m32), linker scripts** and ran on QEMU.

BroLang Compiler + Custom Virtual Machine (Java & JVM-style Execution Model) [GitHub]

- Designed a complete compiler from scratch for a custom language (BroLang), implementing a **Lexer, Parser, AST, and Bytecode Generator**.
- Built a **custom 16-bit Virtual Machine (VCPU + 65KB RAM)** to execute compiled bytecode, inspired by **Java + JVM-style execution**.
- Developed a full runtime system supporting **instruction decoding, stack-based execution, control flow, and I/O operations**.
- Engineered a self-contained execution ecosystem enabling high-level programs to run on custom virtual hardware.
- **Tech Stack:** Modern C++17, custom ISA design, bytecode architecture, manual memory management.

OS-Level AI Agent: Autonomous System Controller via LLM Tool Execution [GitHub]

- Built an **OS-level AI agent** that converts any LLM API into a **fully autonomous system controller** capable of executing real-world tasks.
- Designed a **planner → action → feedback loop** enabling the agent to perform multi-step reasoning and tool-based execution.
- Implemented system tools including **shell command execution, file I/O, web interaction, and Python runtime**, enabling full OS control.
- Engineered **self-healing behavior** where the agent detects failures (e.g., missing dependencies) and autonomously resolves them (e.g., installing packages via pip).
- Integrated optional browser automation (Playwright) and built a **GUI-based control interface** for real-time task execution.

Technical Skills

Languages: C++, Python, TypeScript/JavaScript, SQL
AI & Distributed Training: LLM Pretraining & Fine-tuning, TPU (v4/v6e) & GPU (A100/H100) Clusters, JAX **pmap**, Scaling Laws, Mixed Precision (BF16), XAI (SHAP)
Frameworks & Web: JAX/Flax, PyTorch, HuggingFace, React, Next.js, FastAPI, PostgreSQL, MongoDB
Cloud & Systems Architecture: GCP (TPU Research Cloud), AWS, Docker, Linux, Compiler Design, Virtual Machines, POSIX Sockets

Achievements

Awarded Google TRC grant for 320 TPU access (SPOT + STANDARD VMs).
Awarded AMD MI300X GPU access for 300 non-preemptive hours
IOQM (PRMO): Qualified for Indian Olympiad Qualifier in Mathematics [Certificate]
KVPY SA: AIR 2590 [Certificate]
LeetCode: Knight (2013), 300+ problems solved [Profile]
CodeChef: 4-Star (1818) [Profile]
Codeforces: Specialist (1417) [Profile]
CodeChef START166B: Global Rank 9 [Contest]