Rudransh Shekhar

Prayagraj, UP, India | +91 8429632553 | kanhaiffco2007@gmail.com
Portfolio: Rudransh-Shekhar | **mySphere**: mysphere.co.in | **KaptaanLM Labs**: kaptaanlm.tech
GitHub: github.com/rudy-07 | **LinkedIn**: linkedin.com/in/rudransh-shekhar
Hugging Face: huggingface.co/kaptaan45 | **Kaggle**: kaggle.com/kaptaan45

Profile

Computer Science undergraduate at Vellore Institute of Technology (VIT), Chennai (**CGPA: 9.44, Branch Rank 3**), Founder at **KaptaanLM Research Labs** and **mySphere**, and former AI/ML Intern at the **India Meteorological Department (IMD)**. Specializing in the architecture and distributed pre-training of high-parameter-density small language models (SLMs), local LLM C++ inference runtimes, scalable distributed cloud architectures, and autonomous multi-agent consensus networks. Experienced across distributed TPU/GPU clusters, large-scale foundation pre-training datasets, and high-performance computing (HPC).

Education

Vellore Institute of Technology (VIT) — Chennai, India
Bachelor of Technology in Computer Science and Engineering; ***CGPA: 9.44*** *(Branch Rank 3)* — *2025 – 2029*

R.N. Public School (CBSE Board) — Prayagraj, India
Senior Secondary Education (Class XII: ***91.6%*** *| Class XI:* ***91.1%****)* — *2024 – 2025*

St. Joseph's College (ICSE Board) — Prayagraj, India
Secondary Education (Class X: ***96.8%*** *|* ***100/100 in Computer Applications****)* — *2023*

Experience & Technical Research

KaptaanLM Research Labs (kaptaanlm.tech) — Remote
Founder & Independent AI Researcher — *Aug 2026 – Present*

- **KaptaanLM Fleet & Scout-10M Pre-training**: Architected a modular family of compact SLMs (Scout 10M, Vanguard 25M, Navigator 50M, Commander 75M, Sovereign 100M → MoE). Pre-trained baseline Scout-10M across 5.0 Billion tokens on Google TPU pods using custom 16k BPE tokenization, tied embeddings, QK-Norm, and dual logit soft-capping to resolve small-model repetition collapse.
- **QaptaanLM-0.75B Causal LM**: Stripped Alibaba's Qwen3.5-0.8B vision transformer down to a lean 752M causal LM; orchestrated a two-stage training lifecycle (1B tokens Continued Pre-Training on KapCode-1B and 100M tokens Supervised Fine-Tuning on KapInstruct-100M) on Google TPU v5e-8.
- **InferenceOS Adaptive Engine**: Designed and engineered an open-source, hardware-aware local LLM operating system featuring a universal multi-format engine (GGUF, ONNX, PyTorch, OpenVINO), Kahneman System 1 & 2 symbiosis, dynamic microbatching, live hysteresis layer migration, and dual OpenAI/Ollama REST servers.
- **Open Dataset Suite & HF Ecosystem**: Curated and published a 1.1B+ token dataset suite (KapCode-1B, KapInstruct-100M) in Apache Arrow and Parquet formats (50% FIM infilling, assistant loss masking); released across 8 Hugging Face repos with turnkey GGUF (13 quantizations), BitsAndBytes 4-bit NF4, and ONNX WebGPU runtimes.

mySphere (Udyam Registered MSME | mysphere.co.in) — Remote
Founder & Lead Systems Architect — *Aug 2025 – Present*

- **myCloud Distributed Personal Cloud**: Architected an enterprise personal cloud decoupling metadata orchestration from distributed storage nodes; developed live WebSockets relay, CLI tool (`mycloudctl`), Python SDK (`mycloud-sdk`), Desktop Node GUI, and Oracle Cloud (OCI) backend.
- **myTrade Multi-Agent Platform** (Product): Engineered an autonomous market intelligence platform for Indian equities (NSE/BSE), pairing quantitative indicators with a 40+ LLM jury consensus mechanism, ChromaDB vector RAG, friction tax modeling, and remote Telegram HITL execution.
- **myDomain Dynamic Cloud Platform**: Engineered developer dynamic DNS management integrating Cloudflare APIs and DDNS; built automated DNS record synchronization (A, AAAA, CNAME, TXT, MX) with 1-click cloud platform presets and sub-domain health checks.
- **myCam, mySync & myTune Ecosystem**: Engineered low-latency WebRTC IP surveillance streaming with ray-casting polygonal activity zones (myCam); built cross-device WebRTC P2P sync with on-device Llama 3 summaries (mySync); developed dual-platform Electron and web audio streaming suite with Web Audio API multi-band EQ (myTune).

Experience & Technical Research (Continued)

India Meteorological Department (IMD), Ministry of Earth Sciences New Delhi, India
AI/ML Intern & Research Collaborator *Aug 2025 – June 2026*

- **Pangu-Weather & Pangu-Lite HPC Scaling**: Fine-tuned Huawei's Pangu-Weather on 3.5 years of IMD GFS atmospheric data (2023–2026), slashing epoch training from ∼20 to ∼8 minutes; scaled lightweight Pangu-Lite on 40+ years of ERA5 data, reducing epoch training time from ∼32 hours to ∼5 hours via DDP and NCCL on IMD's Arunika HPC cluster.
- **Model Conversion & Validation**: Reverse-engineered open-source Pangu-Weather checkpoints; transformed ONNX models into PyTorch (`.pth`) weights with verified numerical parity and benchmarked inference consistency across frameworks.
- **Auto-Regressive Forecasting Pipelines**: Developed auto-regressive multi-step forecasting pipelines generating weather predictions up to 10 days ahead; built preprocessing, validation, and conversion pipelines for GRIB, GRIB2, and NetCDF meteorological formats.
- **Agentic AI Assistant & Cluster Workflows**: Developed an agentic RAG weather assistant integrating local/cloud LLMs (Zephyr, Gemma) with Mausamgram datasets; authored PBS job scheduling scripts (`qsub`, `qstat`), checkpointing mechanisms, and geospatial evaluation workflows across multi-node Linux Red Hat cluster environments.

Open-Source AI Contributions & Community Releases

Hugging Face Transformers (Merged PR #49960 | PR #49960 | 168k★) Python, PyTorch, Audio ML

- Diagnosed and resolved Issue #23648 in `ClapFeatureExtractor`: fixed boolean `padding=True` evaluation failure in mel extraction that erroneously caused audio inputs to fall through to constant zero padding instead of the configured `self.padding` strategy (`repeatpad`); added regression test suite `test_padding_boolean` to upstream CI.

Career-Ops AI Agent & Latent Spaces Brag Video Skill (Merged PRs #5822, #15 | 78k★, 19k★) Agent Workflows

- Engineered structured CLI argument routing for `--help` and `-h` in Career-Ops autonomous job search agent; integrated native Google Antigravity (AGY) agent skill discovery pathways and workspace asset resolution in Latent Spaces Brag.

Technical Skills & Tools

AI/ML & Foundation Models: Small Language Models (SLMs), Continued Pre-Training (CPT), Supervised Fine-Tuning (SFT), PyTorch, JAX/Flax, Distributed Training (DDP, NCCL), llama.cpp, RAG, ChromaDB, Agentic AI, Quantization (GGUF, BitsAndBytes NF4), ONNX WebGPU, MediaPipe, NLP.

Programming Languages & Systems: Python, C / C++, Java, Node.js / Express, Flask, SQL, Shell/Bash Scripting, R, MATLAB, REST APIs, WebSockets (Socket.IO), WebRTC P2P Mesh.

Cloud, HPC & Infrastructure: Google TPU v5e, NVIDIA CUDA, High-Performance Computing (HPC), PBS Job Scheduling (`qsub`), Oracle Cloud (OCI), Cloudflare DNS, Linux (Red Hat, Ubuntu), Nginx, Gunicorn, systemd, Docker, Git/GitHub, Hugging Face Hub, PyPI.

Data Systems & Scientific Formats: Apache Arrow, Parquet, SHA-256 Deduplication, MySQL, SQLite, ChromaDB Vector DB, GRIB, GRIB2, NetCDF.

Frontend & App Development: React 19, React Native, TypeScript, Vite, CustomTkinter, Electron, Capacitor, HTML5, CSS3, JavaScript, HTML5 Canvas API.

Honours & Awards

Certificate of Merit – Branch Rank 3 (2026): Awarded for securing 3rd rank in CSE branch at Vellore Institute of Technology (VIT), Chennai; presented by Mr. Binod Kumar, MD & CEO of Indian Bank.

Academic Excellence Scholarships (2025, 2023): Awarded Rs. 3,28,000 scholarship by IFFCO for 91.6% in CBSE Class 12th (2025), and Rs. 24,000 for 96.8% in ICSE Class 10th (2023).

Certificate of Excellence in Computer Applications (2023): Trophy and certificate from St. Joseph's College for scoring a perfect 100/100 in Computer Applications in ICSE Board Examinations.

State Academic Media Features (2023, 2025): Recognized in regional newspaper coverage (Amar Ujala) for 96.8% (ICSE) and 91.6% (CBSE), and named in VITEEE toppers rankings.

Certifications & Languages

Certifications: Infosys Springboard (Advanced Features in Java) | ECMWF (Machine Learning for Earth Systems Modelling, Climate Data Analysis) | NASA ARSET (Remote Sensing & Hyperspectral Data) | EUMETSAT (Satellite Observation & H SAF Products).

Languages: English (Fluent / Full Professional), Hindi (Native), German (Elementary).