

SIDDHARTH N R

AI / MLOps Engineer & Founder

Tamil Nadu, India · +91 99946 13811 · siddhufx7@gmail.com

github.com/SiddharthFX · huggingface.co/Pluto-AI-Labs · linkedin.com/in/siddharth-n-r-842529356

PROFESSIONAL SUMMARY

AI Engineer and Founder of Pluto AI Labs, building open-source models with 9,000+ cumulative Hugging Face downloads and contributing to 8+ open-source AI organizations, including Apple's MLX. Specializes in distributed MLOps, QLoRA fine-tuning, and fault-tolerant training on constrained hardware, with two peer-reviewed publications and a growing research focus on LLM verification systems.

TECHNICAL SKILLS

- **MLOps & Training:** Distributed Data Parallel (DDP), QLoRA, Cross-Session Checkpoint Recovery, Hugging Face Hub API, Weights & Biases, Docker, GitHub Actions, CI/CD
- **Frameworks:** PyTorch, Hugging Face Transformers, PEFT, TRL, bitsandbytes, scikit-learn, Imms-eval
- **Model Optimization:** Knowledge Distillation, 4-bit NF4 Quantization, Apple MLX Conversion & Quantization, GGUF/llama.cpp, MinHash Deduplication
- **Verification Research:** NLI, Embedding-Based Relation Scoring, LLM-as-Judge Evaluation, Synthetic Data Validation
- **Web & Cloud:** Python, FastAPI, React, Node.js, Google Cloud Run, Supabase, MongoDB Atlas

EXPERIENCE

Crebiliti Limited

AI/ML Research Contributor

Sep 2026 – Present

England, UK (Remote)

- Conduct AI/ML research on evidence positioning for LLM systems, combining deterministic attribute engines with model-based inference to raise verification accuracy from 78% to 98% on controlled benchmarks.
- Design and execute controlled experiments across claim-evidence relation categories (NLI, embeddings, LLM judges), producing an attribute-level failure matrix that guides the research roadmap.
- Build a validated synthetic-data pipeline for AFU classifier training, combining local LLM generation, dual-NLI validation, and leakage-tested dataset splits with full provenance.

Pluto AI Labs

Founder & Lead AI Engineer

Jan 2026 – Present

Remote, India

- Founded an independent AI research lab; shipped open-source edge models generating 9,000+ cumulative Hugging Face downloads and became an active contributor to 8+ open-source AI organizations, including Apple's MLX, where released work includes quantized MLX-format models.
- Architected a 68-hour 2x T4 DDP QLoRA pipeline for Apollo-VL-Edge-3B, a 3B vision-language model matching Qwen2.5-VL-3B-Instruct on ChartQA (78.6%) and AI2D (78.0%) while outperforming larger 4B+ models – 300+ downloads in 48 hours.
- Curated and released Athena-VL-250K, a 250,000-row vision-language corpus powering Athena 1.0-4B, engineered to deliver 9B-class document and chart intelligence on consumer hardware.
- Fine-tuned Pluto-Genesis-0.6B via QLoRA to a 44.02% benchmark average (#4 globally among sub-1B models, 2,000+ downloads in 48 hours), solving a HuggingFace Trainer failure mode with a 3-layer cross-session checkpoint recovery system.
- Fine-tuned Atlas-Coder-2-0.5B past Qwen2.5-Coder-0.5B and DeepSeek-Coder-1.3B on strict EvalPlus benchmarks (36.6% HumanEval+); packaged as a 400MB GGUF running at 75 tokens/sec on 8GB MacBooks.

Codec Technologies

Blockchain Developer (Intern)

Jan 2026 – Mar 2026

Remote, India

- Designed and implemented ERC-1155 multi-token smart contracts on Ethereum with a distributed New Zealand and India team, applying secure infrastructure design that directly informed the peer-reviewed Sentient Audit research.
- Developed Solidity smart contracts for Ethereum applications, building deep familiarity with code-level vulnerability patterns.

MegaBliss Worldwide <i>AI-Based Blockchain Developer (Intern)</i>	Aug 2025 – Nov 2025 <i>Remote, New Zealand</i>
Built real-estate NFT infrastructure on Ethereum with AI-assisted contract logic and automated compliance for an international client, delivering ERC-1155 multi-token smart contracts with a distributed New Zealand and India team.	
Igress Solutions LLP <i>Machine Learning Engineer (Intern)</i>	Jun 2024 <i>Remote, Bangalore, India</i>
Developed LLM-integrated machine learning pipelines for data analysis and automated insight generation, and built Power BI dashboards for client-facing reporting.	

PROJECTS

llm-diff • <i>PyPI CLI, “git diff for LLM behavior” (Flagship)</i>	Jun – Jul 2026
Behavioral regression testing for LLMs: backend-agnostic client (Ollama, OpenAI, Groq), regex-based scorers across behavioral dimensions, and CI/CD-ready exit codes for GitHub Actions and Jenkins.	
AuditLink AI • <i>Smart Contract Audit Platform</i>	Apr – May 2025
Real-time, on-chain audits of verified Solidity contracts: Chainlink Functions fetch source from Etherscan, Gemini scores vulnerabilities, and verified results are minted as NFT badges for transparency.	
AnvilX • <i>Blockchain Desktop App + Web Platform</i>	
End-to-end macOS and web platform built solo: authentication flows, on-chain functionality, and production-ready deployment pipelines across the full stack.	
Crypfolio • <i>Privacy-First Crypto Tracker</i>	Jun – Jul 2025
Browser-only crypto portfolio tracker with live CoinGecko market data and P&L analytics; zero-account, local-only persistence that auto-expires after 48 hours.	
MedASK AI • <i>Healthcare AI Platform</i>	
OCR-to-insight pipeline that reads prescriptions via Google Cloud Vision and uses Gemini to return plain-language clinical guidance, deployed end-to-end on Google Cloud Run.	

PUBLICATIONS

- Siddharth N R, “Pluto-Genesis: Reliable QLoRA Fine-Tuning on Ephemeral Compute with Cross-Session Checkpoint Recovery for Sub-1B Language Models,” *Zenodo (CERN)*, July 2026. zenodo.org/records/21368749
- Siddharth N R, “Sentient Audit: An AI-Driven Multi-Pass Security Analysis Platform for Smart Contract Vulnerability Detection and NFT Certification,” *IRJMETs*, 2026.

EDUCATION

Ramco Institute of Technology (Anna University) <i>B.Tech, Artificial Intelligence and Data Science</i>	2022 – 2026 <i>Tamil Nadu, India CGPA: 7.3/10</i>
Coursework: Machine Learning, Deep Learning, NLP, Computer Vision, Neural Networks, Data Structures and Algorithms	
P.A.C.M Higher Secondary School <i>Higher Secondary (Class XII), Computer Science</i>	2022 <i>Tamil Nadu, India 78.5%</i>
P.A.C.M Higher Secondary School <i>Secondary (Class X)</i>	2020 <i>Tamil Nadu, India 78.2%</i>

CERTIFICATIONS AND LANGUAGES

- Certifications:** AI Capabilities and Limitations (Anthropic); Claude 101 (Anthropic); Machine Learning Specialization, DeepLearning.AI/Coursera (Andrew Ng); Full Stack AI Development (Google Antigravity and Cursor AI)
- Languages:** English (B1, Professional); Tamil (Native); Telugu (Mother Tongue); German (A1, In Progress)