NIRANJAN KRISHNA

+91 9188 473 712 | niranjankrishna.acad@gmail.com
linkedin.com/in/theniru | github.com/niranjanorkat

Python & ML Engineer with 5+ years of experience building scalable data and ML pipelines. Specialized in training and fine-tuning large language models, optimizing inference, and automating foundation model evaluation for production-grade performance.

EXPERIENCE

FULLSTACK ENGINEER | AUG 2025 — CURRENT

VA Labs, 6omb

- Built a platform for multi-tenancy voice agents to help bring 33% increased sales to customers with tele-calling in the US and EU.
- Increased code quality by 60% while refactoring incumbent platform built in React, Node and Firebase.
- Piloted database schema standardization, to reduce incumbent attributes by 12%, reducing bugs and increasing system reliability and robustness.
- Led development of Super Admin, as an emulation layer on top of existing application, improving RBAC and ensuring 20% increased admin efficiency.

APPLICATION ENGINEER | NOV 2022 — APR 2025

Formant

- Led development of the core fleet management portal, increased Formant's total revenue by ~30%.
- Piloted portal with 99.8% uptime for managing 20,000+ production robots.
- Enhanced CSAT to >92% by leading customer success engineering, writing internal documentation, and engineering core libraries & pipelines in Python, Go, gRPC, & ROS2 for real-time robot control and communication
- Reduced latency from 15 minutes to <1 second in data ingestion for processing 500,000+ records per month by building fault-tolerant, distributed data ingestion pipelines.
- Processing 1M+ data points in real time with minimal latency for actionable operational insights by developed performance analytics tools using Snowflake and Python.

AI ENGINEER | NOV 2021 — APR 2022

Reknow.ai

- Saved over $200k in cost for chatbot applications by fine-tuning GPT-J language models locally in Python.
- Automated deployment of large language models using Docker, reducing setup time by over 70% and enabling faster experimentation cycles.
- Achieved 85% accurate resolution of user queries by developing clustering-based QA models in PyTorch for user query resolution.

LEAD SOFTWARE ENGINEER | JAN 2021 — JUL 2021

FindMonster

- Created AR gameplay experiences using Niantic ARKit, incorporating real-world environmental awareness.
- Achieved IoU score > 0.7 by implementing semantic segmentation models to identify and classify natural objects for accurate AR object placement.

LEAD SOFTWARE ENGINEER | JAN 2020 — DEC 2020

TheGGLife

- Supported avg 150k+ concurrent users in live streaming by engineering live streaming server architecture in Node.js via WebSockets and integrating Kafka for reliable event streaming and backpressure handling.
- Reduced debugging and incident response times by 20–30% through Prometheus-based monitoring and alerting setup for live server environments.
- Achieved 94% classification accuracy in command-to-gameplay translation via NLP using PyTorch in Unity-based multiplayer games with live-streams.

SKILLS & EXPERTISE

- Python
- PyTorch
- TensorFlow
- scikit-learn
- Hugging Face (Transformers)
- LangChain
- OpenAI API
- NumPy
- Pandas
- FAISS
- ONNX
- Docker
- Kubernetes
- FastAPI
- Weights & Biases (wandb)
- AWS (EC2, S3)
- GCP (Vertex AI)
- CI/CD Automation
- Model Evaluation & Monitoring
- Vector Databases (Chroma, FAISS)
- Git
- Snowflake

PUBLICATIONS

"Classier Guided Diffusion for Image Inpainting. Applications to Fine Art", Accepted at **LXAI at ICML 2022**