Swaraj Purohit

LinkedIn: linkedin.com/in/swarajpurohit
GitHub: github.com/anomius

Email: swarajpurohit@gmail.com
Mobile: +91 74058 26741

Professional Summary

Senior AI Engineer with **4+ years** of production experience in **LLM/VLM systems**, **agentic architectures**, and **multimodal AI**. Expert in designing, fine-tuning, and deploying **production-ready GenAI applications** using **PyTorch, Hugging Face Transformers**, and **agentic frameworks** (LangChain, LlamaIndex) serving **10k+ users**. Strong background in **transformer architectures**, **parameter-efficient fine-tuning** (LoRA/QLoRA), **quantization**, and **distributed training**. Interdisciplinary researcher with expertise in **quantum computing** and **quantum machine learning** (Delft University QuTech, NYUAD mentor), combining **neurosymbolic methodologies** with **reinforcement learning** for advanced AI systems. Published **quantum optimization framework to PyPI** and led research on **QAOA algorithms** for real-world applications. Proficient in **Python** and **C++** for high-performance AI deployment.

Core Technical Skills

- **LLM/VLM & Agentic AI**: **Large Language Models** (GPT-4, Claude, Llama), **Vision-Language Models**, **Transformers**, **Azure OpenAI/OpenAI APIs**, **Hugging Face**, **LangChain**, **LlamaIndex**, AutoGPT, CrewAI, **RAG Architectures**, Multi-agent coordination, Tool orchestration, Memory systems
- **Fine-tuning & Optimization**: **LoRA/QLoRA**, Adapter Layers, Prefix Tuning, **Quantization** (4/8-bit, GPTQ), **Model Distillation**, RLHF, RLAIF, Prompt Engineering, **DeepSpeed**, vLLM, TensorRT, ONNX Runtime
- **Quantum Computing**: **Qiskit**, **QAOA**, **QUBO modeling**, **Quantum graph coloring**, Variational Quantum Circuits, Quantum-classical hybrid algorithms, **Quantum ML**, **RL for quantum systems**, IBM Quantum hardware, Published **qsavvy** to PyPI
- **ML/DL Frameworks**: **PyTorch** (Expert), **TensorFlow**, Scikit-learn, CNNs, RNNs, LSTMs, GANs, **Distributed Training**, Model Parallelism, OpenCV
- **Infrastructure & Tools**: **Python/C++/Rust**, **Pinecone, FAISS**, Weaviate, Milvus, Redis, Kafka, **Docker**, Kubernetes, AWS/Azure/GCP, CUDA, FastAPI, **Langfuse**, Weights & Biases, MLflow

Professional Experience

- **Deloitte US–India Offices** — Bangalore, India
  AI & Data Analyst (LLM/GenAI Specialist) — *Jan 2024 – Present*
  - **Production LLM Systems**: Co-developed enterprise **RAG prototypes** using **OpenAI API**, **LangChain/LlamaIndex**, and **Pinecone** for Fortune 500 clients. Architected **agentic workflows** with multi-step reasoning, tool use orchestration, and memory systems. Implemented **evaluation frameworks** measuring answer faithfulness, contextual precision, and response quality.
  - **Model Fine-tuning & Optimization**: Experimented with **LoRA** and **adapter-based fine-tuning** for domain-specific LLM adaptation. Implemented **quantization strategies** (4-bit, 8-bit) reducing model size by **75%** while maintaining 95% accuracy. Optimized **embeddings** and chunking strategies, reducing retrieval latency by **30%**.
  - **Distributed ML Pipeline Engineering**: Led MLOps for pharmaceutical forecasting platform across **20 new markets**. Built **distributed training pipelines** with **Kubernetes** and **Argo CD**. Implemented automated retraining, model versioning, and A/B testing infrastructure. Reduced deployment time by **70%**.
  - **Scalable Data Engineering**: Architected cloud-native ETL pipelines using **PySpark** and **AWS/Azure** processing 35+ data sources. Optimized batch processing with distributed computing, reducing latency from 8+ hours to under 5 minutes while improving reliability by **20%**.
- **Vedika.ai** — Remote
  Senior AI Engineer - LLM/Agentic Systems (Contract) — *2023 – 2024*
  - **Multi-tenant RAG Platform**: Architected scalable **agentic RAG platform** using **Azure OpenAI**, **LangChain**, and **Pinecone** serving **10k+ MAU**. Built autonomous agent frameworks with **reasoning loops**, **tool use orchestration**, and **memory systems**. Implemented **multi-agent coordination** for complex query decomposition and parallel processing.
  - **Advanced LLM Engineering**: Developed comprehensive **prompt engineering** frameworks with chain-of-thought reasoning, few-shot learning, and task decomposition. Built **LLM evaluation harness** with automated testing measuring **F1, exact-match, BERTScore, answer faithfulness**. Implemented **RLHF-style feedback loops** for continuous model improvement.
  - **Production Inference Systems**: Built low-latency inference systems using **FastAPI** microservices, containerized with **Docker** and deployed on **Azure**. Implemented **efficient inference pipelines** with response caching, model routing, and **distributed serving**. Reduced **p95 latency by 40%** and token costs by **30%**.
  - **Multimodal AI Integration**: Integrated **vision-language models** for document understanding and visual question answering. Built hybrid retrieval systems combining sparse (BM25) and dense (embeddings) retrieval, reducing hallucinations by **25%**.
- **Quidich Innovation Labs** — Mumbai, India
  Computer Vision & Deep Learning Engineer — *2022 – 2023*
  - **Real-time Vision Systems**: Developed production **CNN**-based tracking systems using **PyTorch** and **TensorRT** deployed in **ICC WTC, IPL, T20 World Cup**. Achieved sub-50ms latency with 99% tracking accuracy. Implemented **transformer-based architectures** (HRNet) for pose estimation.
  - **Edge AI & GPU Optimization**: Optimized models for **edge deployment** on Jetson platforms using **TensorRT** and **ONNX Runtime**. Applied **quantization** (INT8, FP16) and model distillation, reducing inference time by **25%** while maintaining production accuracy. Implemented **CUDA kernels** for custom operations.
  - **Automated ML Infrastructure**: Built distributed training pipeline with **ClearML** on AWS. Implemented automated hyperparameter tuning, data augmentation, and model versioning. Deployed **CI/CD pipelines** for continuous model retraining and evaluation.

- **Delft University of Technology, QuTech** — Netherlands (Remote)
  Quantum Computing Research Intern — *Jul 2021 – Aug 2021*
  - **Quantum Reinforcement Learning Research**: Developed **reinforcement learning agents** for quantum information theory experiments under Dr. Aritra Sarkar (QuTech). Applied **universal AGI RL framework** to model quantum observables and optimize quantum measurement protocols, improving measurement accuracy by **4%** through adaptive learning strategies.
  - **Neurosymbolic AI for Quantum Systems**: Explored **neurosymbolic methodologies** combining quantum computing with classical machine learning. Investigated hybrid quantum-classical architectures bridging symbolic reasoning with quantum state manipulation, contributing to foundations of quantum-enhanced AI systems.
  - **Competitive Achievement**: Selected as one of **30 international participants** from 1000+ global applicants in QWorld Association's **QIntern 2021 program**. Achieved **3rd place** in quantum computing competition among 50 teams, demonstrating expertise in quantum algorithm design and implementation.

Quantum Computing Research & Open Source

- **qsavvy: Quantum Computing Framework (2022 – Present)**: Author and maintainer of **cross-platform quantum computing toolkit** published to **PyPI** with 1000+ downloads. Developed **8 quantum optimization algorithms** including quantum graph coloring, **quantum annealing**, and **variational quantum circuits** showing **4% improvement** over classical methods. Implemented **quantum circuit transpilation** across multiple backends (IBM, Rigetti, Amazon Braket) with comprehensive documentation and testing. PyPI — Docs — GitHub
- **NaQa: Quantum Graph Coloring (2023)**: Developed **novel quantum graph coloring algorithm** with **4% quantum advantage**. **QUBO encoding** with **QAOA on IBM Quantum**. Hybrid system with **ARIMA ML models** for UAE agriculture optimization. addressing UAE agriculture challenges. for market prediction, optimizing multi-year crop distribution across soil quality, environmental factors, and market demand parameters. GitHub
- **Quantum Algorithm Research**: Researched **quantum-classical hybrid algorithms** and **variational quantum circuits** for combinatorial optimization. Explored applications in graph problems, scheduling, and resource allocation. Investigated **quantum machine learning** architectures combining quantum feature maps with classical neural networks, demonstrating measurable quantum advantage in specific problem domains.

Research Publications

- **GAN-Based Plant Disease Detection (2023)**: "Augmenting Training Data with Generative Adversarial Networks." Achieved **98% accuracy** using novel GAN-augmented CNN with contrastive loss. Demonstrated expertise in **deep generative models** and training optimization. Springer Nature
- **Satellite Data Reconstruction with GANs (2021)**: "Reconstruction of Missing Data in Satellite Imagery Using SN-GANs." Applied spectral normalization and Wasserstein loss for stable GAN training in remote sensing applications. Springer Nature

Key Projects & Open Source Contributions

- **NuShell Core Contributor (2024)**: Active contributor to **NuShell** (26k+ stars), modern shell written in **Rust**. Merged production PRs deployed in releases v0.101.0, v0.105.1, v0.106.0 impacting thousands of daily users. Demonstrates **C++/Rust** proficiency for performance-critical system components and large-scale codebase navigation.
- **Hermes.AI: Agentic Chatbot Platform (2022)**: Built scalable **multi-agent chatbot platform** supporting 112+ languages with **transformer-based** knowledge retrieval. Implemented **reasoning loops** and tool use orchestration. **Winner: Smart India Hackathon 2022**. Pioneered agentic workflows combining symbolic reasoning with neural architectures before LangChain/AutoGPT popularization.
- **WebGL Neural Network Accelerator (2022)**: Developed browser-based CNN accelerator using **WebGL/WASM** achieving **98% speed improvement** over JavaScript. Implemented custom **GPU kernels** for convolution operations, demonstrating expertise in **low-level optimization** and edge device deployment strategies.

Education

- **Dr. Vishwanath Karad MIT World Peace University** — Pune, India
  B.Tech. in Computer Science & Engineering; GPA: 8.95/10 — *2019 – 2023*
- Relevant Coursework: Deep Learning, Reinforcement Learning, Quantum Computing, Computer Vision, Distributed Systems

Certifications & Recognition

- **NVIDIA Generative AI with LLMs (2025)**: Associate Certification - Advanced LLM techniques
- **Google Cloud AI/ML Stack**: Vertex AI, BigQuery ML, GenAI Studio Professional
- **Qiskit Global Summer School (2023)**: Quantum Machine Learning Specialization - Selected from 1000+ applicants
- **IBM Neuro-Symbolic AI (2023)**: Professional Certificate in Advanced AI Systems
- **Smart India Hackathon Winner (2022)**: National-level AI competition - MHRD India
- **QCHACK Stanford & Yale Winner (2021)**: International quantum computing hackathon - Innovative quantum algorithms
- **NYUAD Hackathon Mentor (2023)**: Technical mentorship in quantum computing & optimization