Yuvraj Garg

yuvraj97.ml@gmail.com +91 9351967450 Bengaluru, India quantml.org

github.com/yuvraj97 linkedin.com/in/yuvraj97

SUMMARY

Machine Learning Engineer with 4+ years of experience architecting end-to-end production AI systems. Proven track record of independently engineering scalable AI products and leading a 3-person R&D team from concept to deployment. Specialized in computer vision pipelines, multi-agent orchestration (LLMs/ RAG), and MLOps.

PROFESSIONAL EXPERIENCE

Machine Learning Engineer, *Styldod* 10/2021 – Present | Bengaluru, India

- Lead a 3 person AI/ ML R&D team building computer vision, 3D, and generative AI solutions for visual automation.
- Architected a 2D to 3D scene reconstruction pipeline using an ensemble of ML models and linear algebra for spatially accurate 3D furniture placement from single images.
- Launched REimagineHome's first Image AI Agent, utilizing an MCP architecture and LangChain to orchestrate a multi-agent system featuring inter-agent communication. This collaborative architecture automated complex tool calling across image generation APIs, database fetching, live Amazon API integrations, and content moderation.
- Engineered an automated content pipeline targeting AI search engines, driving a 17% increase in new user acquisition. Architected agentic tool calling to handle autonomous internet research, cross linking, and citation generation.
- Cut operational costs by 70% by architecting a production MLOps system that deploys self-hosted LLMs and image generation models, serving millions of monthly requests.
- Slashed VLM and image generation cold start latencies by over 85% (down to 50s and 37s) by optimizing inference pipelines, maximizing API throughput and enabling efficient GPU auto-scaling.
- Trained Diffusion models across multi GPU clusters on millions of images. Built a custom Human-in-the-Loop evaluation pipeline to bypass unreliable loss graphs and ensure high fidelity production deployments.

Research Engineer, *RAx Labs* 07/2021 – 10/2021 | Gandhinagar, India

- Developed an NLP pipeline and Twitter bot for real-time research paper summarization, enhancing algorithms for sentence ranking and information extraction.
- Deployed scalable backend architectures via Docker on AWS ECS with a full observability stack (Prometheus, Grafana, DataDog) for production monitoring.
- Optimized heavy Elasticsearch queries indexing thousands of complex PDFs, minimizing retrieval latency for academic research data.

Skills

- **Core Competencies:** Leading AI/ML R&D Teams , End-to-End Product Ownership , Product Strategy & Roadmapping , Growth Engineering.
- **AI & Machine Learning:** Computer Vision & 3D Understanding , Generative AI , Large Language Models (LLMs) , Vision-Language Models (VLMs) , Agentic Orchestration , LangChain/LangGraph , AI Agents & MCP (Model Context Protocol) , Statistical Modeling.
- **Engineering & MLOps:** Production AI Deployments, Edge AI & Local Inference, MLOps Training/Deployment at Scale, Serverless Architectures, Data Visualization.
- **Languages & Frameworks:** Python , PyTorch , FastAPI, React & Next.js , Ray.
- **Cloud & Infrastructure:** AWS Cloud Architecture , Google Cloud Platform (GCP) , Linux & Docker

Production AI Deployments

ScaleWaveAI, *Serverless AI Inference Infrastructure* 2022 – 2025

- Architected a serverless AI inference platform with auto scaling GPU clusters, global CDN deployment, and a credit based billing system.
- Served 15+ production computer vision models (including segmentation, depth estimation, and zero shot classification) via unified REST APIs, enabling instant deployment without manual setup.

KyoudaiClub, *LLM Character Fidelity with Real-Time PvP* 2025 – 2026

- Launched a full stack AI platform for real time character roleplay using Next.js and FastAPI, engineered for low latency streaming.
- Implemented advanced LLM context management to maintain long term character consistency across extended user sessions.
- Built a live WebSocket driven quiz engine featuring dynamic question generation, HP/combo mechanics, and real time PvP matchmaking.

QuantML, *Making Machine Learning Interactive* 2021 – Present

- Built an interactive ML education platform featuring live visual twins of algorithms, allowing users to inspect weight matrices and step through real training loops.
- Shipped comprehensive lesson modules combining visual narratives, live model inspectors, PyTorch code, and context aware AI assistants.
- Engineered edge deployment pipelines hosting multimodal LLMs Qwen3.5 0.8B (Text/Image) and Gemma 4 E2B (Text/Image/Audio) optimized for local inference on standard 8GB RAM desktops.

Overheard, *WebSocket-Driven Live Audience Chat Platform* Early 2026

- Built a live public chat platform using a WebSocket broadcast architecture to deliver low latency real time conversations to concurrent audience viewers.

CERTIFICATES

- **Deep Learning & Statistics:** Fundamentals of Statistics | Probability - The Science of Uncertainty and Data | Machine Learning with Python | Data Analysis in Social Science | Neural Networks and Deep Learning | Convolutional Neural Networks | Improving Deep Neural Networks | Structuring Machine Learning Projects .
- **Infrastructure & Automation:** Red Hat Certified System Administrator (EX200) | Red Hat Certified Specialist in Ansible Automation | Red Hat Certified Specialist in OpenShift Administration

EDUCATION

MicroMasters in Statistics and Data Science, edX (Online)
Massachusetts Institute of Technology (MITx) via edX
Advanced Machine Learning, Probability, Data Analysis, and Statistical Modeling.