Rakesh Nagaraju

Santa Clara, CA | +1 (669) 288-4508 | rakesh.nju2205@gmail.com | linkedin.com/in/Rakesh-Nagaraju|
github.com/Rakesh-Nagaraju | https://www.rakesh-ai.com

SUMMARY

AI Engineer with 5+ years of experience in AI research and software development, specializing in vision, LLMs, deep learning, and MLOps.

Proven track record in training, finetuning, optimizing, monitoring, deploying and leading teams to build scalable, innovative AI solutions.

TECHNICAL SKILLS

- **Expertise**: Python, Computer Vision, LLM Engineering, RAG Systems, AI Agents, Model Optimization
- **Frameworks & Libraries**: PyTorch, TensorFlow, HuggingFace, LangChain, Langraph, LlamaIndex, YOLO (v3/v4/v8), FastAPI, NumPy, Pandas
- **MLOps & Infrastructure**: AWS (EC2, S3, SageMaker, Lambda), VectorDB, Docker, Kubernetes, GitLab CI/CD, MLflow, Langfuse, Wandb.
- **Development & Testing**: Git/GitHub API, pytest, REST APIs, Object-Oriented Design, JSON Schema Validation
- **Visualization & Front End**: Chainlit, Streamlit, Gradio, React, HTML/CSS, JavaScript

EXPERIENCE

AI Engineer – Uniquify Inc - Santa Clara, USA — **Aug 2021 – Present**

Software AI Agent with GitHub Integration - LlamaIndex, MCP, httpx, Codellama, Github API, Langfuse

- Engineering a modular AI agent that acts as a software developer—writing and completing code, managing repos via MCP server tools.
- Developed error-resilient GitHub integration—automated branch creation, issue triage, merge conflict resolution—boosting merge success from 70% to 95%.
- Instrumented real-time prompt→response tracing in Langfuse, cutting mean time to debug failures by 40%.
- Contributing to ongoing development of additional modules and testing suites to extend the agent's functionality and robustness.

RAG-Based Chatbot for Internal SoC Documents - MinIO, Milvus, LlamaIndex, Chainlit

- Led hybrid search RAG chatbot serving 100+ daily users with 90%+ answer relevance by fine-tuning models on SoC domain queries.
- Implemented RLHF feedback loop by fine-tuning a reward model on user preferences using PPO—improving user satisfaction.
- Deployed a fully containerized solution with a real-time Chainlit front end and CI/CD-driven model promotions
- Integrated fallback mechanisms and additional retrieval models to ensure high relevance and low latency in production, while using Langfuse to monitor LLM outputs and leveraging judge templates to assess data quality, consistency, and output integrity.

Person and Object Detection on Embedded Devices - YOLOv3/v8, CNN, AWS

- Led neural network team to fine-tune and quantize CNN, Detectron, and YOLO models for face analysis and object detection on embedded devices.
- Converted models to 8-bit ONNX, achieving 2.5× throughput with >95% accuracy across real-world conditions.
- Integrated models into CI/CD pipelines with firmware teams, ensuring smooth production deployment.
- Developed Dockerized annotation UI with PostgreSQL, enabling users to correct predictions and support continuous re-training.

Comprehensive AI Training Curriculum - NLP, CV, LLMs

- Authored end-to-end curriculum on transformers, vision, and LLM pre-training; onboarded 50+ engineers with 90% course-completion.
- Created hands-on modules and tutorials for advanced AI concepts project-ready proficiency by 60% (capstone scores).
- Mentored 15+ junior engineers through complex AI challenges, leading to their successful transition into industry roles.

Defect Detection System - YOLOV4_Tiny, GitLab CI/CD

- Fine-tuned YOLOv4-Tiny for defect detection, boosting accuracy from 85% to 98% and cutting inspection costs by 50%.
- Established pytest-based automated testing pipelines in GitLab CI/CD, achieving zero production regressions.

Senior Software Engineer – Capgemini - Bengaluru, India — **Nov 2016 – Aug 2019**

Backend software development and maintenance - python, DB2, beautifulsoup, OOP

- Developed and maintained Python applications to extract tax-related data from DB2 databases.
- Performed data operations to align with business needs, created APIs for access, and deployed them.
- Supported User Acceptance Testing (UAT) for successful software deployment.
- Implemented object-oriented programming principles and optimized code for performance.

EDUCATION

MS in Computer Science, San Jose State University, San Jose, CA — **Aug 2019 - May 2021**

- **Coursework:** Machine Learning, Artificial Intelligence, Distributed Computing, Cybersecurity

PUBLICATION

Published a paper titled 'Generating Fake Malware Using Auxiliary-Classifier GAN for Malware Analysis' (arXiv:2107.01620, 2021).