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 computer vision, deep learning, and MLOps. Skilled in deploying models to production, optimizing performance, and leading technical teams to deliver innovative solutions.

TECHNICAL SKILLS

- **Expertise:** Python, Computer Vision, LLM Engineering, RAG Systems, AI Agents, Model Optimization
- **Frameworks & Libraries:** PyTorch, TensorFlow, LlamaIndex, HuggingFace, YOLO (v3/v4/v8), FastAPI, NumPy, Pandas
- **MLOps & Tools:** AWS (EC2, S3, SageMaker, Lambda), GitLab CI/CD, Docker, MLflow, Langfuse, Milvus, MinIO
- **Development:** Git/GitHub API, pytest, Object-Oriented Programming, JSON Schema Validation
- **UI/Visualization:** Chainlit, Streamlit, Gradio, HTML, CSS, JavaScript, ReactJS

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.

· Integrated mcp sever with GitHub specific tools for branch management, issue managements, and robust error handling.

· Monitoring performance, prompt quality, and agent behavior with Langfuse for real-time observability and root-cause debugging

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

· Led an end-to-end RAG chatbot project for internal SoC documentation, combining hybrid dense–sparse retrieval

· Fine-tuned the Gemma reranker on domain-specific query–document pairs to boost relevance, and tracked experiments

· Implemented an RLHF loop to continuously refine the LLM's responses: collected user feedback, trained a reward model, and performed policy optimization—logging each iteration in MLflow for auditability

· Deployed a fully containerized solution with a real-time Chainlit front end and CI/CD-driven model promotions, complete with MLflow Model Registry integration for both reranker and LLM artifacts

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

· Led end-to-end development of person and object detection on embedded devices, integrating YOLOv3/v8 for object localization and DeepFace for high-accuracy facial recognition across varied lighting and pose conditions

· Quantized both YOLO and DeepFace models to 8-bit precision and converted them to ONNX format via our custom export pipelines, ensuring seamless compatibility with our proprietary SoC runtime SDK

· Validated ONNX models on the target hardware—measuring real-world inference latency, throughput, and power consumption—and automated CI/CD firmware integration with MLflow-logged performance metrics for reproducibility and continuous monitoring

Comprehensive AI Training Curriculum - NLP, CV, LLMs

- Developed extensive training materials covering a wide spectrum of AI topics including QA, transformers, vision, and LLM pre-training.
- Created hands-on modules and tutorials for advanced AI concepts, facilitating understanding of both foundational algorithms and state-of-the-art techniques.
- Mentored 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 manufacturing defect detection achieving 98% accuracy and 50% cost reduction
- Implemented automated testing pipelines with pytest and GitLab CI/CD for robust deployment

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).