Srikar Madarapu

(470) 738-4285 | Sunnyvale, California | srikarmadarapu24@gmail.com
www.linkedin.com/in/srikar8 | Srikar8.github.io

SUMMARY

AI Engineer with 7+ years of experience designing and deploying production-grade AI systems across Enterprise, startups, and research. Specialized in LLM-powered **agentic workflows, SDLC automation, computer vision,** and **multimodal AI**. Successfully delivered AI first products from 0→1 into production, integrating with enterprise systems while ensuring reliability, cost efficiency, security, and scalability. Comfortable taking **end-to-end** ownership of system architecture and translating complex business workflows into robust AI solutions.

SKILLS

Agentic & LLM Systems: LLMs, RAG, Lang Chain, Lang Graph, Lang Smith, Langfuse, LiteLLM, True Lens, Crew AI, OpenAI Agent Builder, Chat Kit, Prompt Engineering, FAISS, Pinecone, Milvus, Temporal Workflows, Hugging Face, Co-Pilot
Programming & Frameworks: Python, TypeScript, Node.js, Fast API, React, PostgreSQL, MongoDB, Linux, REST/GraphQL APIs
Machine Learning & AI: PyTorch, Transformers, CLIP, Vision Transformers, CNNs, YOLO, Semantic & Instance Segmentation, SLAM, Few-shot Learning, Transfer Learning, Hyper-parameter Tuning, Scikit-learn, Pandas, NumPy, JAX
MLOps & Deployment: Docker, Kubernetes, CI/CD, Jenkins, Airflow, Argo CD, MLflow, Weights & Biases, Metaflow, ONNX, Model Optimization
Cloud & Enterprise Platforms: AWS (S3, SageMaker, Lambda, EC2, Bedrock, Kafka), Azure (AI Foundry, Document Intelligence, Blob Storage), Snowflake, Databricks

EXPERIENCE

Forward Deployed Solutions Engineer - Applied AI **Mar 2026–Present**
Cisco (Contract) *Sunnyvale, California*

- Owned requirements discovery, prototyping, and rollout of agentic automation solutions for Nexus NX-OS engineering teams in Data Center Networking, translating operational bottlenecks and developer feedback into reusable agent capabilities.
- Built CI investigation and remediation agents that correlate Jenkins logs, test results, and code changes to distinguish product defects, flaky tests, and infrastructure failures, generating fixes with human approval.
- Reduced AI workflow latency by **40%** and cost by **60%**; implemented semantic log filtering and pattern classification to reserve LLM reasoning for cases requiring deeper analysis.
- Cut regression test execution time by **50%, from 4 hours to 2 hours**, through regression analysis and test workflow optimization, accelerating validation of NX-OS changes.
- Built shared agent infrastructure and a unified RAG system powering Webex bots and engineering workflows, incorporating source citations, metadata filtering, and retrieval evaluation.
- Collaborated across NX-OS, Catalyst IOS XE, and IOS XR teams to surface infrastructure failures obscured by passing tests, deliver diagnostic reports, and recommend integration fixes; integrated pyATS/Genie checks to validate device state before and after approved changes.
- Deployed and scaled Langfuse and built agent evaluation harnesses to support troubleshooting and iterative reliability improvements; contributed to AI-assisted code review and SDLC automation spanning bug triage, fixes, testing, and merge workflows.
- Supported developer adoption of the internal AI platform through onboarding, tool guidance, and workflow integration, feeding engineering needs back into platform improvements.

Full-Stack AI Engineer **Jan 2025–Mar 2026**
Right Pro Inc *Sunnyvale, California*

- Led end-to-end development of an **AI-powered procurement platform**, reducing manual operations by 85%.
- Architected **agentic AI systems** for autonomous multistep planning and execution using structured tool calling, persistent memory, secure action schemas, and policy-based guardrails.
- Built production-grade conversational AI systems with **workflow** orchestration and state management.
- Applied prompt engineering, tool-augmented reasoning, retrieval-augmented generation (RAG) to improve accuracy and reliability.
- Integrated AI agents with REST/Graph QL APIs, databases, document pipelines, and internal services.
- Implemented **MCP servers** and a decorator-based **tool registry** for automatic API discovery and agent–tool integration.
- Built secure autonomous email and document processing agents using Microsoft Graph API and Azure AI Document Intelligence.
- Implemented a secure, centralized, **multi-tenant** document storage and retrieval layer on Azure Blob Storage.
- Containerized services and workflows using Docker, deploying via Azure DevOps and Argo CD with automated CI/CD pipelines.
- Implemented **observability** and **monitoring** (logging, metrics, tracing) to support production SLAs and rapid incident response.
- Drove **0→1** AI product development, rapidly prototyping and scaling using modern AI coding tools (Cursor, Claude code).
- Collaborated with product, design, and **customers** to translate complex business workflows into scalable AI solutions.
- Ensured security, compliance, and data privacy across AI pipelines, including secrets management and access isolation.
- Took ownership of system design, architecture, and production rollout, mentoring engineers where applicable

Computer Vision Research Assistant **Aug 2023 – Nov 2024**
MedIX Lab DePaul University *Chicago, Illinois*

- Implemented an advanced semantic image search pipeline, training **OpenAI CLIP** architecture from scratch and optimizing FAISS vector database for sub-second retrieval Link.
- Developed a medical CT image **multi-classification** system for lung cancer, **fine-tuning** state-of-the-art neural networks to achieve 8% improvement in diagnostic accuracy.
- Reduced model training time by 75% through **MLflow** tracking, hyperparameter optimization, and efficient pipeline design.
- Conducted comparative analysis of **Vision Transformers** versus **CNNs** for medical image classification, identifying specific architectures that excel with limited medical datasets.
- Implemented a parameterized CNN model for image color enhancement based on published research, achieving 89% similarity to target image quality metrics.

Graduate Teaching Assistant **Apr 2024 – Jun 2024**
DePaul University *Chicago, Illinois*

- Provided academic support to 60+ students, simplifying complex concepts and assisted with coursework/projects related to data science, machine learning and image processing.

AI & Visual Data Analytics Founder **Sep 2019 – Dec 2022**
SRP *Bangalore, India*

- Established and grew a startup in **visual media production**. Cultivated strong relationships with agencies, vendors, freelancers, and creative teams. Successfully completed 20+ projects with 6+ partnered companies and traveled across India.
- Developed **financial dashboards** in Tableau to track revenue, expenses, and project performance.
- Leveraged **CVAT** for annotating image/video datasets, automated data extraction via API.
- Implemented **ETL** pipelines to transform annotations for seamless ML integration.
- Developed aerial and 360° image and **video Data Warehouses** for AI-driven projects.
- Led digital marketing initiatives and analyzed customer behavior, driving a 20% increase in conversions.
- Conducted A/B testing and analytics, optimizing SEO and digital ad performance.

Video Analytics Engineer **May 2019 – Aug 2019**
Origa Labs *Bangalore, India*

- Designed real-time motion detection system for human activity monitoring, enhancing machinery safety alerts by 35%.
- Engineered a multi-object detection and tracking system for industrial safety, integrating Kalman filters and Deep SORT algorithms to maintain object identity through occlusions.
- Optimized model inference with Open VINO, achieving 3x speed improvement on edge devices through quantization and pruning.
- Developed containerized deployment pipeline with Docker, creating reproducible environments for vision system.

Computer Vision Engineer **May 2018 – Sep 2018**
Inferigence Quotient LLP *Bangalore, India*

- Conducted experiments on various open-source SLAM systems using camera and IMU data.
- Identified localization challenges for indoor map and improved accuracy by 5%.
- Calibrated cameras to ensure accurate localization and 3D mapping in indoor.

Research Intern **Jul 2017 – Dec 2017**
DRDO CAIR *Bangalore, India*

- Implemented CNN-based depth prediction from monocular images based on research paper and achieved 78% accuracy.
- Integrated depth prediction model with SLAM system for improving real-time robotic navigation.

EDUCATION

Master's in Data Science, *DePaul University, Chicago* Jan 2023 – Nov 2024
Bachelor's in Computer Science, *NIIT University, India* July 2014 – May 2018

CERTIFICATION

AI Code Review - DeepLearning.AI
AI Agents for Image and Video Generation - DeepLearning.AI
AWS Certified Generative AI Developer
AWS Certified AI Practitioner
Oracle Cloud Generative AI Professional
Long-Term Agentic Memory with Lang Graph
Building Ambient Agents with Lang Graph