Devi Eswar Kumar Damerla

devieswar79@gmail.com | linkedin.com/in/devi-eswar-kumar-damerla | devieswar.github.io

Professional Summary

Full-Stack Software Engineer with 5+ years of experience delivering scalable, production-grade systems across web, cloud, and AI/ML domains. Expert in full-stack development leveraging FastAPI, Node.js, NestJS, Angular, and React with proven capability to architect and deploy end-to-end applications. Specialized in embedding AI/ML solutions including Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and knowledge graph architectures into production environments. Proficient in cloud infrastructure (AWS SageMaker, Bedrock, ECS Fargate), microservices architecture, and DevOps methodologies (Docker, Kubernetes, CI/CD). Demonstrated track record of generating measurable business impact through technical innovation and cross-functional collaboration.

Professional Experience

AI Engineer Nov 2024 – Present
AI for ALL LLC (AI OWL) *Columbus, OH*

- Lead engineer for Flash AI, a CJIS-compliant evidence intelligence platform enabling secure multimedia (video, audio, document) analysis for law enforcement using RAG and multi-agent LLM systems (Claude 3, Legal-BERT).
- Architected and implemented multi-agent orchestration layer using MCP (Model Context Protocol) with JSON-RPC/SSE protocols, enabling agent-to-agent communication for collaborative task execution and escalation handling.
- Designed transformer fine-tuning pipelines using LoRA and QLoRA for legal domain adaptation, boosting semantic retrieval precision by 35% while slashing inference costs by 40%.
- Integrated Weaviate vector database with AWS SageMaker and FastAPI microservices, delivering sub-100ms hybrid search with metadata filtering and role-based access control across 100K+ document collections.
- Orchestrated asynchronous multimodal ingestion pipelines using WhisperX, OpenCV, and FFmpeg for audio-video transcription and scene understanding, accelerating processing time by 60% and enabling Vision-Language Model capabilities.
- Established MLOps pipelines with Docker, ECS Fargate, and SageMaker for continuous model evaluation, versioning, model governance, and audit compliance in production environments.
- Mentored junior engineers on RAG architecture design, LLM prompt engineering, and vector database schema optimization; contributed to AI product roadmap and strategic integration initiatives.

Software Engineer Aug 2021 – Jul 2023
Airbus Group *Bengaluru, India*

- Enhanced aircraft 2D/3D CAD model loading performance by 75% through microservices-based iPaaS application leveraging NestJS, Angular, AWS Application Load Balancer, and S3 optimization strategies.
- Delivered MEAN stack application that decreased CATIA license dependency by 60%, cutting annual software costs by $120K+ while maintaining design quality through WebSocket-based model streaming.
- Deployed OAuth-based Single Sign-On (SSO) and comprehensive API documentation using Swagger/OpenAPI; participated in Agile/Scrum and SAFe ceremonies across full SDLC.
- Automated design quality checks and optimized SQL queries from 3+ minutes to under 60 seconds, accelerating data processing for 200+ engineering workflows.
- Configured CI/CD pipelines using Jenkins on AWS ECS Docker containers with SonarQube integration for code quality gates; executed smooth MariaDB-to-MySQL migrations with zero downtime.

Junior Data Scientist Aug 2019 – Jun 2021
MCR Web Solutions *Andhra Pradesh, India*

- Architected optimized CNN models for image denoising, deblurring, and low-light enhancement using TensorFlow, achieving 83% structural similarity (SSIM) improvement through hyperparameter tuning and data augmentation.
- Engineered automated ML pipeline integrating preprocessing, training, and inference with Python, TensorFlow, and AWS Lambda, cutting processing latency by 60% and enabling on-demand image enhancement at scale.
- Deployed trained models as Docker-based microservices via RESTful APIs on AWS infrastructure for serverless execution with auto-scaling compute allocation.
- Performed exploratory data analysis (EDA) on large-scale image datasets using NumPy, Pandas, and Matplotlib to identify noise patterns and optimize feature extraction strategies.

Machine Learning Engineer Intern May 2019 – Jul 2019
TheSmartBridge Private Limited *Hyderabad, India*

- Created machine learning model for chronic kidney disease diagnosis achieving 95% accuracy using Naive Bayes, decision trees, and random forest algorithms; deployed as Django web application with user-friendly interface.

Education

Master of Engineering in Computer Science Apr 2025
University of Cincinnati, College of Engineering and Applied Science *Cincinnati, OH*

- GPA: 3.67/4.0
- Relevant Coursework: Machine Learning, Deep Learning, Computer Vision, Natural Language Processing, Cloud Computing, Software Engineering

Projects

ArthaNethra — AI Financial Risk Investigator | *Python, FastAPI, Neo4j, Weaviate, Angular* Oct 2024 – Nov 2024

- **Financial AI Hackathon Finalist** — Engineered knowledge graph-native financial investigation platform that automatically maps entities and relationships across multiple documents (10-Ks, contracts, invoices) into interactive graph database.
- Designed dual-database system using Neo4j for entity relationships and Weaviate for semantic vector search, enabling complex cross-document queries impossible with traditional RAG approaches.
- Constructed hybrid AI extraction pipeline combining LandingAI ADE for structured data and AWS Bedrock Claude for narrative parsing, achieving 99% accuracy at 80% lower cost than pure LLM approaches.
- Built explainable AI chatbot with clickable PDF citations and automatic graph generation; integrated Sigma.js for interactive visualization with multiple layout algorithms.
- Deployed full-stack application using Docker Compose; slashes M&A due diligence time from 200+ hours to under 2 hours (90% time reduction).

Context-Driven Image Narration | *React, Node.js, Python, TensorFlow, Transformers* Sep 2023 – Feb 2024

- Led development of MERN stack application with deep learning model achieving 79.7% accuracy in object recognition using TensorFlow and Hugging Face Transformers for context-based image captioning and narration.
- Built data preprocessing pipeline for large-scale datasets and improved model performance through transfer learning with pre-trained vision-language models.

Technical Skills

Programming Languages: Python, JavaScript, TypeScript, Java, Go, SQL, C#

AI/ML Technologies: LLMs, RAG, Knowledge Graphs, Multi-Agent Systems, LoRA/QLoRA, Prompt Engineering, PyTorch, TensorFlow, scikit-learn, Hugging Face, LangChain, LlamaIndex, Pandas, NumPy, OpenCV

Databases: Neo4j, Weaviate, Pinecone, PostgreSQL, MySQL, MongoDB, Redis, Elasticsearch

Cloud & DevOps: AWS (SageMaker, Bedrock, Lambda, ECS Fargate, S3), Docker, Kubernetes, Jenkins, CI/CD, Git

Frameworks: FastAPI, Flask, Django, Node.js, NestJS, Angular, React, Spring Boot, ASP.NET Core

Development Practices: Agile/Scrum, Microservices, RESTful APIs, TDD, MLOps, Model Governance

Certifications

AWS Certified Solutions Architect – Associate
Deep Learning Specialization — DeepLearning.AI (Coursera)