

TARUN DEEP REDDY B V

Los Angeles, CA 90007

+1 (213) 210-1143 tarundeepreddybv@gmail.com [LinkedIn](#) [GitHub](#)

Summary

AI/ML Engineer and Researcher with hands-on experience in Machine Learning, Deep Learning, Computer Vision, NLP, and Generative AI. Experienced in building RAG pipelines, AI agents, and real-time vision systems, with a focus on model optimization and evaluation. Author of a research paper accepted for publication at ICCP 2026.

Education

University of Southern California (USC)

M.S. Computer Science (General)

Los Angeles, CA

Aug 2026 – May 2028

SRM Institute of Science & Technology

B.Tech Computer Science & Engineering, CGPA: 8.7/10

Chennai, India

Aug 2022 – May 2026

Experience

AI/ML Student Researcher

April 2025 – May 2026

SRM Institute of Science & Technology

Chennai, India

- Accelerated real-time facial recognition inference 12x to ~1.1s and reduced CPU utilization to 23% through optimized Haar Cascade/LBPH computer vision pipelines.
- Reduced demographic subgroup performance variance by ~18% using Context-Aware Synthetic Data Generation (CSDG), Selective Fine-Tuning (SFT), and adaptive loss reweighting.
- Evaluated model robustness across demographic subgroups and challenging conditions including noise, blur, and face-mask occlusion; authored a research paper accepted for publication at ICCP 2026.

Projects

Multimodal Autonomous Agentic AI System

2026

- Built a multimodal pipeline using OpenCV, Vision Transformers, and Whisper to process text, image, video, audio, and PDF inputs into unified embeddings.
- Developed agentic AI + RAG with LangGraph, LlamaIndex, and FAISS for planning, tool calling, memory, self-reflection, RAGAS evaluation, and FastAPI/Docker deployment.

SKYNET: Silicon Valley Job Monitoring AI

2026

- Built an AI job intelligence pipeline using Python/NLP to process LinkedIn, Telegram, and Instagram postings and rank opportunities by technical skill relevance.
- Developed domain-weighted job matching, ATS resume generation, document tracking, face-recognition authentication, voice interaction, and application analytics.

AI Powered Workplace Monitoring System

2025 – 2026

- Developed a real-time computer vision system using Haar Cascade, LBPH, and MediaPipe for workplace presence and attendance monitoring.
- Implemented CSDG, SFT, adaptive loss reweighting, an 80% temporal validation rule, and robustness testing under noise, blur, and face-mask occlusion.

Skills

Languages: Python, C, C++, Java, SQL

ML/AI: Machine Learning, Deep Learning, Computer Vision, NLP, Generative AI, LLMs, RAG, Agentic AI, Multimodal AI, Fine-Tuning, Model Evaluation, Model Optimization

Frameworks: PyTorch, TensorFlow, Transformers, OpenCV, MediaPipe, LangGraph, LangChain, LlamaIndex, Whisper, FastAPI, Scikit-learn

Tools & Infrastructure: FAISS, ChromaDB, MLflow, NumPy, Pandas, PostgreSQL, SQLite, AWS, Docker, CUDA, REST APIs, Git/GitHub, MCP

Certifications

- **Machine Learning Specialization** — DeepLearning.AI & Stanford Online
- **Deep Learning Specialization** — DeepLearning.AI
- **Agentic AI** — DeepLearning.AI
- **SQL for Data Science with Python** — IBM