

SRAVANAKUMAR SATHISH

Chandler, AZ (Open to Relocate) | +1 (480) 743-1232 | sathishsravanakumar@gmail.com | F-1 OPT (STEM Eligible)
[linkedin.com/in/sravanakumarsathish](https://www.linkedin.com/in/sravanakumarsathish) | github.com/sathishsravanakumar | [sravanakumar-sathish-portfolio](#)

SUMMARY

AI/ML Engineer shipping production agentic AI, RAG, and computer-vision systems across enterprise (PwC) and applied AI startups. Deep hands-on expertise in LangChain, PyTorch, FastAPI, and cloud-native deployment (AWS, Azure, GCP). M.S. Data Science candidate at ASU, STEM OPT eligible.

EDUCATION

Arizona State University, Tempe - M.S. Data Science, Analytics and Engineering Aug 2024 - May 2026
Vellore Institute of Technology (VIT), Chennai - B.Tech Electronic and Computer Engineering Jul 2019 - Jul 2023

PROFESSIONAL WORK EXPERIENCE

AI Engineer (Part-Time) - qBotica, Inc., Phoenix, AZ Jul 2026 - Present

- Built a **LiveKit** voice agent (web + phone) with **HubSpot CRM** caller verification that automates end-to-end freight order booking into **Degama dtmsWeb** via a three-tier pricing engine (contract tariff, mileage estimate, human fallback).
- Eliminated hallucinated-price checkout errors for BayOps AI's multi-vendor auto-parts ordering platform by adding price-sanity validation and **Twilio/Playwright/Claude Computer Use** cart automation, with live billing synced to Excel via **xlwings**.
- Improved chatbot response accuracy 30% over semantic and fixed-size chunking baselines by developing an agentic-chunking **RAG/NLP** pipeline with context-aware document segmentation and retrieval tuning.

AI/ML Developer (Part-Time) - DriverAI, Peoria, AZ (Remote) Jun 2026 - Present

- Achieved 80%+ model accuracy training and validating **computer-vision** models on live camera-feed data, processing 200,000+ images through an **AWS S3**-based ingestion and training pipeline for production-scale baselines.
- Selected **YOLOv12n** for production deployment after benchmarking four **YOLO** architectures (v8n, v12n, v12s, v26n) on **mAP**, GPU memory footprint, and inference latency, optimizing for edge-inference constraints over marginal accuracy gains.

AI/ML Engineer (Capstone) - Honeywell Aerospace, Tempe, AZ Jan 2026 - May 2026

- Delivered a live Mesa Fire Department pilot of **FirstWatch**, an AI-powered wearable mental-health platform (smartwatch, mobile app, command dashboard), leading a 5-person team through architecture and deployment.
- Projected to cut burnout 30% and accelerate critical intervention 40% by designing a real-time biometric data pipeline and **predictive ML models** for stress and crisis detection.

Generative AI Associate Consultant - PwC, Bangalore, India Jan 2023 - Jun 2024

- Cut contract-analysis runtime 70% for enterprise legal workflows by building production AI microservices with **FastAPI** and **Azure Document Intelligence** for automated contract parsing and clause extraction.
- Accelerated semantic search and record retrieval 60% across large-scale document ingestion by engineering HyreGPT, an enterprise RAG/LLM pipeline using **LangChain**, **FAISS** vector search, and **Azure OpenAI**.
- Reduced inference latency 40% and manual-review overhead 48% by optimizing model inference and data pipelines through **PyTorch** quantization, while preserving data-privacy compliance.

ML Engineer Intern - Atsuya Technologies, Chennai, India Jul 2022 - Dec 2022

- Achieved 100% uptime with zero cloud dependency by building a real-time **YOLO** object-detection system on Android (**TensorFlow Lite**) for Indian Oil Corporation's industrial safety compliance, backed by an offline-first **Kotlin** inference framework for low-connectivity environments.

PROJECTS

RACE - RAG-Optimized Clinical Reasoning Engine Dec 2025 - Feb 2026

- Compressed **Llama-3-8B** from 32 GB to 5.5 GB by fine-tuning with **QLoRA** (rank-16 adapters, 0.1% of 8B parameters) and 4-bit **NF4 quantization**, enabling deployment on consumer 8 GB VRAM GPUs.
- Mitigated LLM hallucinations by architecting a compound AI pipeline integrating **RAG** with **ChromaDB** vector retrieval and **LangChain** orchestration, grounding every clinical response in verified medical evidence.
- Achieved sub-second evidence retrieval latency by building an end-to-end pipeline using **sentence-transformers** (all-MiniLM-L6-v2) for semantic embedding and Top-K similarity search from a persistent vector store.

Pitch Deck Debater - Multi-Agent Startup Critique System Oct 2025 - Jan 2026

- Simulated realistic investor boardrooms by orchestrating a **Multi-Agent System** with specialized AI personas (VC, CTO, Product Lead) to autonomously critique startup pitch decks.
- Improved pitch resilience metrics ~40% by designing an adversarial feedback loop utilizing **Chain-of-Thought** prompting to generate counterarguments and strategic pushback.
- Resolved inter-agent hallucinations via **consensus-voting logic** by implementing stateful conversation management that synthesizes conflicting agent perspectives into a cohesive strategic report.

ChronosGuru - Persona-Driven Historical RAG Platform Sep 2025 - Nov 2025

- Achieved 92% factual accuracy for historical-persona conversations by engineering a **RAG** architecture anchored to a curated, domain-specific vector database, with dynamic prompting sustaining persona fidelity across 20+ conversation turns.
- Cut retrieval latency under 200ms and boosted session engagement 3x by integrating **hybrid keyword + semantic search** into the retrieval pipeline.

SKILLS

Generative AI & Agentic Systems: Multi-Agent Orchestration (LangGraph, LangChain), Model Context Protocol (MCP), RAG, Model Evaluation, Voice Agents, Fine-Tuning (LoRA/QLoRA), Transformer Architectures, LLMs, SLM, Vision-Language Models, Prompt Engineering, Playwright, Computer-Use Automation
Machine Learning: PyTorch, TensorFlow, Computer Vision, NLP, Vector Search & Embeddings (FAISS, ChromaDB), MLOps, Causal Inference, Databricks
Engineering & Architecture: Python, FastAPI, Docker, Microservices, System Design (reliability, scalability, fault-tolerance patterns), Git/GitHub, CI/CD Pipelines, Node.js, React.js
Cloud & Data Platforms: AWS (S3), Azure (AI Studio, Azure OpenAI, DevOps), GCP (BigQuery, Vertex AI), SQL, PostgreSQL, Streamlit, Power BI

CERTIFICATIONS AND PUBLICATIONS

Certifications: Azure AI Engineer Associate (AI 102), Advanced MLOps (Databricks), MCPs: Advanced Topics (Anthropic)
Publication: "Music Recommendation via Facial & Speech Emotion Recognition," IEEE ICETET 2024