SUMAN MADIPEDDI

madipeddisuman@gmail.com| (602) 565-9192 | linkedin.com/in/suman-madipeddi | github.com/SumanMadipeddi

SUMMARY

AI Engineer with **2+ years** of experience building production-grade systems in Generative AI, LLMs, and Computer Vision. Skilled in deploying scalable ML pipelines, fine-tuning, and optimizing inference. Passionate about applying research to real-world AI products.

EDUCATION

M.S. Robotics and Autonomous Systems **May 2025**
Arizona State University, Tempe, Arizona 3.74 GPA

- **Relevant Coursework:** Artificial Intelligence, Applied AI and ML, Reinforcement Learning and Perception in Robotics

TECHNICAL SKILLS

Deep Learning & Vision: PyTorch, TensorFlow, Keras, CNNs, RNNs, GANs, Computer Vision, Multimodals (CLIP, BLIP, Flamingo)

AI/ML & GenAI: LLMs (GPT, Gemini, Llama 3, Claude, BERT), Transformers, VLMs, HuggingFace, LangChain, LlamaIndex, RAG, FAISS, Pinecone, Conversational Agents, Fine-tuning, RLHF, Diffusion Models, OSS ML models, JAX, AGI, OpenAI, Twilio

Languages & Frameworks: Python, C, C++, Go, Swift, Java, JavaScript, TypeScript, Node.js, FastAPI, React Native, CUDA

Data & MLOps: Pandas, NumPy, ETL, SQL, GPU, BigQuery, PostgreSQL, Databricks, Amazon (SageMaker, Bedrock), CI/CD, kafka

Cloud & DevOps: AWS (S3, EC2, Lambda), GCP, Azure, Docker, Linux, Snowflake, Terraform, Agile, Apache Spark, GitHub, REST APIs

PROFESSIONAL EXPERIENCE

Minor Chores, USA: AI and LLMs Engineer **Jan 2025 – Present**

- Designed and deployed a **RAG-based conversational AI system** (Vertex AI, RLHF) with multimodal inputs, achieving 90% intent accuracy, and built an **LLM-based recommendation engine** leveraging geospatial + behavioral data for real-time chore-preneurs
- **Modernized UI/UX across iOS/Android** by leading React Native, Swift, and TypeScript development within a microservices architecture; orchestrated deployment with Kubernetes, ensuring scalable, zero-downtime rollouts.
- **Improved performance and engagement** by reducing latency 40%, introducing broadcast + interactive messaging, and driving measurable gains in onboarding 25% and user engagement 30%.

PROJECTS

Voice AI Agent for Automated Lead Qualification – Setter.AI (Twilio, Deepgram, Typescript, Docker) **Aug 2025**

- **Engineered a Node.js (TypeScript) + Twilio** outbound calling platform, integrating with Go High Level CRM to automate lead engagement within 10 minutes of creation, orchestrated via Docker for scalable reliability.
- **Built a GPT-4 powered conversational AI pipeline** with Deepgram STT/TTS for natural lead screening, appointment scheduling, and real-time CRM data logging through a React/TypeScript dashboard.
- **Delivered 24/7 automation** that eliminated 25+ hours of manual work per week, accelerated response times, and established a scalable foundation for client growth.

Conversational RAG Chatbot Agent with Voice & Video Interface using Google Vertex AI **June 2025**

- Developed a scalable, domain-adaptive AI assistant using **Google Vertex AI** for multiple enterprise clients, designed to deliver accurate, context-aware support from OCR and digital document parsers for internal knowledge bases.
- Built a custom Retrieval-Augmented **Generation (RAG)** pipeline and integrated voice (Google Speech) and **video (Twilio API)** interfaces, enabling seamless multi-modal interaction across **text, voice, and video**.
- Streamlined enterprise support and onboarding workflows by deploying the assistant across varied client environments, improving response relevance and user engagement.

On-Device Real-Time Gesture Recognition System **June 2025**

- Developed a real-time, on-device gesture and voice recognition system that dynamically overlays emojis and confetti effects on live video streams, enhancing user interaction for video calls and streaming platforms.
- Trained gesture classification models with **Flax/JAX**, optimized the inference pipeline with **ONNX Runtime** and **CUDA,** achieving low-latency inference and robust performance across **17** hand gestures. Integrated **Whisper**-based voice command recognition for spontaneous, conversational emoji and effect triggers, supporting multimodal interaction.
- Architected a **modular, privacy-first overlay pipeline** using OpenCV and Pillow enabling integration with **FaceTime and OBS.**

Object Segmentation on ARMBench (PyTorch, R-CNN, ResNet-50) **May 2024**

- Developed a Mask R-CNN model with a ResNet-50 backbone in PyTorch for object segmentation on the ARMBench dataset (50K+ images, 450K+ segments).
- Trained across three subsets (*mix-object*, *zoomed-out*, *same-object*) to evaluate performance in occluded and cluttered scenes.
- Achieved **mAP@50** scores of 0.48, 0.41, 0.48 and **mAP@75** scores of 0.48, 0.06, 0.38 across subsets, using COCO-style evaluation and OpenCV visualizations to monitor segmentation accuracy and generalization and tested on other data ingestion.