

Arjun Sherugar

Orlando, FL | 689-217-5516 | arjunsherugar@gmail.com | U.S. Citizen | [LinkedIn](#) | [GitHub](#)

EDUCATION

University of Central Florida

B.S., Computer Engineering

Orlando, FL

- **Coursework:** Data Structures and Algorithms, Machine Learning, Robot Vision, Object-Oriented Programming

EXPERIENCE

Intrastack Solutions

May 2026 - August 2026

AI & ML Engineer

Orlando, FL

- Shipped customer-facing features across authentication, REST APIs, dashboards, and UI for a production web app in React and Next.js, working inside a 40+ person cross-functional team at an AWS cloud consulting firm.
- Containerized backend services with Docker, ran CI/CD pipelines for automated build, test, and deployment, and provisioned and managed AWS services across development and production environments.
- Led the conversion of the production web app to React Native, architecting shared component logic between web and mobile and implementing mobile-native features to extend the platform to iOS and Android.
- Built proprietary AI/ML features using large language models for enterprise clients.

NASA L'SPACE Mission Concept Academy

January 2026 - April 2026

Command & Data Handling Engineer

Remote

- Owned the Command & Data Handling subsystem design on a 20-member engineering team building a NASA-guided end-to-end mission timeline for a Mars surface expedition.
- Modeled spacecraft geometry and computed entry, descent, and landing trajectories for a NASA Mars landing in Siemens NX and JMARS.
- Conducted CDH trade studies across candidate subsystem architectures, co-developed the mission Software Architecture Flowchart, and integrated cost and schedule data from the Programmatic subteam into the CDH design.
- Co-authored 4 milestone technical documents (MCR, SRR, MDR, PDR) and presented the CDH subsystem design at the Preliminary Design Review, earning a professional-level technical score.

PROJECTS

Tora AI - Local-First Agentic AI Assistant

- Built and shipped a fully local agentic AI assistant that used Qwen2.5 32B via Ollama to drive Gmail, task scheduling, OS and file management, sourced web research, and end-to-end job applications from a phone through a Telegram bot wired into Windows, with 0 cloud API calls in the loop.
- Gated every irreversible action behind human approval, holding 100% of destructive operations pending a phone Approve/Deny card and exposing 0 real secrets to the model, by enforcing deterministic code-based rules over model judgment with asyncio Futures that suspend the agent mid-task and an append-only audit log.
- Automated multi-page ATS job applications end-to-end with zero duplicate submissions and zero tool failures crashing the agent loop, by driving a browser-use agent off live page state instead of CSS selectors, backed by a JSON job store, lenient JSON extraction, repeat-call guards, and error-as-tool-result recovery.

SignBridge AI - Real-Time ASL Accessibility for Video Calls

- Trained and evaluated a 1.2M-parameter Transformer sign classifier to 74% top-1 / 91% top-5 accuracy across 250 ASL signs under held-out model evaluation, using PyTorch on 94,477 MediaPipe landmark sequences and deploying through ONNX Runtime Web at 37–82 FPS.
- Extended the app to unlimited vocabulary at 98% validation accuracy across 28 fingerspelling classes, by training a neural network on 42K hand-landmark images and adding letter-by-letter gesture-based editing.
- Delivered two-way call accessibility with every model running 100% on-device at <1.5 s caption latency, by running streaming Zipformer STT (~22x real-time on CPU) in an isolated Electron utility process and routing Piper neural TTS (~6x real-time) into calls as a virtual microphone with echo-safe gating.

TripWise - AI Travel End-to-End Itinerary Builder

- Shipped a full-stack travel platform covering destination recommendations, streaming AI itineraries, and simulated bookings across 28 async REST/SSE endpoints and 23 automated tests, by building a FastAPI backend over a 10-table PostgreSQL schema with JWT auth against a React 19 and TypeScript frontend.
- Engineered a two-stage recommendation engine and LLM itinerary generation streamed over Server-Sent Events, by blending heuristic scoring (popularity, rating, budget-fit, tag affinity) with pgvector cosine similarity over 384-dim embeddings and applying prompt engineering - schema-enforced JSON outputs, weather-aware prompting, and prompt caching - with per-user Redis rate limiting and per-request token/cost telemetry.
- Cut the main JS bundle 30% via route-level code splitting, by containerizing the stack with multi-stage Docker builds behind an SSE-safe nginx reverse proxy with request-ID tracing and structured JSON logging.

SKILLS

Languages: Python, Java, TypeScript, JavaScript, C, SQL, Kotlin, HTML/CSS

Frameworks & Libraries: React, Next.js, React Native, Electron, FastAPI, Flask, Node.js, PyTorch, ONNX Runtime

Developer Tools & Infrastructure: Git, Docker, AWS, CI/CD, PostgreSQL (pgvector), Redis, nginx, Linux/Bash, pytest

AI/ML: Transformer Training (PyTorch), Model Evaluation, ONNX Deployment, MediaPipe, Embeddings, Vector Search, Vector Databases (pgvector), Prompt Engineering, STT/TTS, Local LLMs (Ollama), Agentic Tool-Calling, Large Language Models (LLMs)

CERTIFICATIONS

- Meta Front-End Programming
- Meta JavaScript
- Duke University Introduction to C Programming