

Sushaanth Srinivasan

(858) 518-9859 | sushaanth.sri@gmail.com | [linkedin.com/in/sushaanth-srinivasan](https://www.linkedin.com/in/sushaanth-srinivasan) | github.com/SushaanthSrinivasan

Software & AI Engineer | Ships end-to-end from research to prod | Fullstack, Agents, LLMs, CV, OCR | ML Masters @ UCSD

EXPERIENCE

Autonomous Machining | AI Engineer

Jul 2025 - Present

Python, Azure, Pytorch, Angular, React, Typescript, LLMs, OCR, AI Agents, RAG, Figma

Seattle, WA

- Built an AI-powered First Article Inspection ballooning tool, outperforming established competitors in customer evaluations
- Fine-tuned custom OCR models for GD&T symbols and manufacturing notation, achieving 95% recognition accuracy, and architected the backend auto-ballooning inference pipeline combining custom-trained and cloud OCR models
- Implemented GD&T callout parsing, default tolerance detection, ASME standard extraction, and automated FAI Form population algorithms, building the core backend logic from drawing ingestion to structured inspection output
- Took full ownership of frontend architecture for a secure document management system with hierarchical navigation, version control, approval workflows, and linked document traceability for regulated manufacturing environments

Pxie Lab, UCSD | ML Researcher

Oct 2024 - Jun 2025

Under Prof. Pengtao Xie | Pytorch, Python, LLMs, Kubernetes, Docker, Selenium, Beautiful Soup

San Diego, CA

- Built InteractGPT, a multimodal LLM (GNN + CNN + Vicuna-13B) trained on 380K+ drug interaction pairs from SMILES
- Achieved 74% macro-F1 on interaction classification, outperforming GPT-4 by 20 points in blinded evaluations
- Built web crawlers to scrape and preprocess 100K+ pharmacological records from 5+ databases

BRG Energies | Software Engineer (Part-time)

Oct 2024 - Jun 2025

Vue, Quasar, Typescript, Taskyon, PGLite, LLMs, RAG

San Diego, CA

- Owned and shipped an agentic AI pipeline, cutting evaluation time by 70% across EPA compliance, zoning, and parcel search

Zemplee | ML Engineer Intern

Jun 2024 - Sep 2024

Flask, Python, React, LLMs, PostgreSQL, Chroma Vector DB, RAG

San Jose, CA

- Cut response time 3x by building a chat copilot to retrieve & analyze sensor data for 20+ healthcare workers & 300+ patients
- Reduced analyst support requests by 40% by building a natural language-to-SQL pipeline with vectorized schemas and LLMs

FincorpX | ML Engineer Intern

May 2023 - Aug 2023

Node, React, Azure OCR, Python, Figma

Chennai, India

- Saved 75% processing time for 50+ small business owners by building an AI financial document platform using Azure OCR
- Achieved 87% text recognition accuracy processing 20+ documents per minute using custom OCR and LLM pipelines

Caterpillar Inc. | SDE Intern

Jan 2022 - Jul 2022

Angular, Javascript, Builder.io

Chennai, India

- Led a 2-intern agile team to adapt Angular UI components for a drag&drop WYSIWYG editor, cutting prototyping time by 66%

EDUCATION

University of California San Diego (UCSD)

San Diego, CA

Master of Science in ECE - Machine Learning & Data Science

Coursework: Recommender Systems, Visual Learning, ML for Physical Applications, Software Engineering, Data Analysis

SSN College of Engineering (Anna University)

Chennai, India

Bachelor of Engineering in Computer Science

Coursework: Internet Programming, Database Management Systems, Data Structures & Algorithms, Operating Systems

TECHNICAL SKILLS

AI/Infra: Python | Pytorch, Tensorflow, Huggingface, CrewAI, Langchain, Vector DB, LLMs | Docker, Kubernetes | Azure

Backend/Data: NodeJs, Flask, FastAPI | SQL, PostgreSQL, NoSQL | Bash | Pandas, NumPy | REST, GraphQL | Spark, Kafka | Git

Web/Mobile: Javascript, Typescript | React/Next, Angular, Vue | HTML, CSS, Tailwind | Selenium | UI/UX, Figma

PROJECTS

Nighthawk - Local-first AI Terminal Autocomplete | *Rust, Ollama, ANSI, Zsh* | [GitHub](#)

- Built an open-source Rust terminal autocomplete with inline ANSI ghost-text, 500 CLI specs & typo-aware token replacement
- Integrated local LLMs via Ollama for context-aware suggestions, with history-based predictions and zero-config setup

Orion - AI browser extension to search websites | *Next/React, Node, Tailwind, Google Gemini, Vector DB, RAG* | [GitHub](#)

- Developed a backend with a custom crawler & graph algorithms to scrape & vectorize subpages, cutting retrieval time by 50%
- Created a chrome extension frontend accessible via a hotkey, enabling intuitive searching across multiple websites