SIDDHANT HITESH MANTRI

Email: siddhantmantri328@gmail.com — LinkedIn: siddhant-mantri — GitHub: siddhant-192 — Phone: (+1) 619-986-7810 — Website: www.siddhantmantri.com

EDUCATION

University of California, San Diego (UCSD) San Diego, CA, USA
Master of Science, Computer Science and Engineering 2025 – 2027
CGPA 3.92/4

Mukesh Patel School of Technology Management and Engineering, NMIMS Mumbai, India
Bachelor's in Technology, Computer Science and Engineering (Cyber Security) 2021 – 2025
CGPA 3.86/4

Indian Institute of Technology, Madras Chennai, India
Bachelor's in Science, Data Science and Applications 2021 – 2025
CGPA 7.00/10, Project CGPA 9.0/10

SKILLS

Programming Languages: Python, Java, C++, JavaScript, SQL, R

ML/AI Frameworks: PyTorch, TensorFlow/Keras, scikit-learn, Hugging Face Transformers, LangChain

Tools & Technologies: Git, Docker, CUDA, Linux, Pandas, NumPy, NLTK, Jenkins, MLflow, Airflow, Selenium, pytest, FastAPI, Node.js

Relevant Coursework: Data Structures and Algorithms, Software Engineering, Database Systems, Computer Networks, Operating Systems, Reinforcement Learning, Deep Learning, Probability and Statistics

Specializations: Large Language Models, Computer Vision, Natural Language Processing, Machine Learning

WORK EXPERIENCE

Qualcomm Jun. 2026 – Present
Machine Learning Intern *Santa Clara, CA, USA*

- Developing Small Language Models (SLMs)

Q-Lab, University of California, San Diego Apr. 2026 – Present
Student Researcher, advised by Prof. Lianhui Qin *San Diego, CA, USA*

- Architecting a model-agnostic system that translates natural-language prompts into physically-realistic, fully populated 3D scenes in Unreal Engine, integrating LLM-based reasoning, retrieval, and geometry-aware generation into a unified framework

Shiley Eye Institute, University of California, San Diego Jan. 2026 – Present
Student Researcher *San Diego, CA, USA*

- Developed a gated attention-based multiple instance learning model fusing fundus, OCT, visual field, and demographic data across **6,195 patients (11,904 eyes)**, achieving **0.90 AUROC** for glaucoma detection, outperforming Transformer (0.87) and late-fusion (0.79) baselines and being clinically deployable
- Designed learned absent-modality tokens enabling robust predictions across **all 31 possible** missing-data combinations; extending framework with temporal modeling to predict glaucoma progression and surgical intervention risk

Computation for Indian Language Technology, IIT Bombay Sep. 2024 – Dec. 2025
Research Intern *Mumbai, India*

- Engineered a pipeline converting Hindi WordNet into **1.25M instruction-response pairs** via dynamic chunking/sliding-window algorithms; fine-tuned **Gemma-3-12B** using **QLoRA 4-bit quantization**, cutting memory usage **75%** (48GB to 12GB)
- Outperformed GPT-4.1 in pedagogical effectiveness (**91.0 vs. 79.4**) with **86% higher response consistency**, validating the system for a Hindi language learning chatbot

Sudha Gopalakrishnan Brain Centre, IIT Madras Jan. 2025 – Jun. 2025
Machine Learning Intern *Chennai, India*

- Trained and enhanced YOLOv9e-seg model on 25-class brain segmentation dataset, achieving **0.97 Dice score** and **30% inference speed improvement** through novel post-processing techniques
- Architected end-to-end ML pipeline using PyTorch for production deployment, processing 1,000+ medical images with automated quality validation and real-time inference capabilities

Volkswagen Group Technology Solutions India Jun. 2024 – Sep. 2024
Machine Learning Intern *Pune, India*

- Improved RAG-based conversational AI system (Project AISHA 2.0) by optimizing vector databases and caching mechanisms, **improving query response time by 40%** and **reducing memory usage by 35%**

PUBLICATIONS

ICTCS 2024 (Springer) 2024

- "Architectural Framework for Automated Incident Response: Leveraging LLMs And Classifiers for Rapid Post-Attack Analysis and Reporting"

NeurIPS 2026 - Under Review 2026

- "SimWorld Studio: Automatic Environment Generation with Evolving Coding Agent for Embodied Agent Learning"