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

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 7.39/10, 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

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

- Engineered structured-data pipeline converting Hindi WordNet into 1.25M diverse instruction-response pairs for specialized conversational AI systems, achieving 91.0 pedagogical effectiveness score vs. 79.4-83.6 for general-purpose models
- Fine-tuned Gemma3-12B language model using LoRA with 4-bit quantization, achieving 58% improvement in response consistency and sub-second inference times while reducing memory requirements by 75%
- Built context-aware keyword detection microservice with FastAPI deployment, improving contextual understanding accuracy by 25% for low-resource language processing

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%

PROJECTS

AI-Powered Learning Management System (LMS) Jun. 2024 – Aug. 2024

- Led the development team implementing AI-powered features, including course summaries, peer-driven insights, and coding assistance, leveraging advanced language models such as LLaMa3-70B to enhance the learning experience
- Integrated the vLLM inference engine to improve the efficiency and performance of LLM queries, reducing latency and scaling AI-generated responses across the system. Implemented FastAPI for serving multiple AI-driven endpoints with Git version control and Scrum development methodology for agile project management, utilizing pytest for unit testing and debugging of application components

ShelfSense - Library Management System Jan. 2024 – Apr. 2024

- Created full-stack web application using Vue.js frontend and Flask backend, implementing comprehensive library operations including book management, user authentication, and issue tracking with role-based access control and RESTful API design, utilizing Selenium for automated testing
- Architected scalable system with SQLAlchemy ORM, Redis caching for improved performance, and Celery task scheduling for automated reminders and report generation using RabbitMQ message broker

PUBLICATIONS

ICTCS 2024 (Springer) 2024

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

IJCNLP-AACL 2025 - Under Review 2025

- "From Lexicon to AI: A Structured-Data Pipeline for Specialized Conversational Systems in Low-Resource Languages"