Sreemukhi Kunche

+91 9676284175 | kunchesreemukhi@gmail.com | Linkedin| Github

Education

Vellore Institute of Technology, B.Tech in Computer Science(AI & ML Specialisation) May, 2026

- **CGPA:** 8.60
- **Coursework:** Artificial Intelligence & Machine Learning, Deep Learning, Applied Machine Learning, Natural Language Processing, Reinforcement Learning, Data Mining, Foundations of Data Science, Probability & Statistics, Linear Algebra, Computer Vision, Data Visualization, DBMS, Operating Systems, Python Programming, AWS Cloud.

Skills

- **Languages:** C++, Python, SQL
- **Machine Learning and AI:** Supervised & Unsupervised Learning, Deep Learning, Neural Networks, NLP, RAG, LLMs, Prompt Engineering, VectorDB, ChromaDB
- **Libraries & Tools:** PyTorch, scikit-learn, NumPy, Pandas, FastAPI, Docker, Git, Langchain

Experience

StealItX | SDE Intern (Remote) December, 2025- Present

- **Built a scalable data extraction pipeline** using Cheerio and Puppeteer to scrape and normalize product metadata across 6+ major e-commerce platforms.
- **Designed an automated monitoring system** using Node-Cron and WhatsApp Web.js to detect price fluctuations and trigger real-time alerts for price drops and restocks.
- **Contributed to live data delivery workflows** by integrating backend services with Firebase Realtime Database for low-latency updates.

Projects

Intelligent LLM Semantic Router *Github*

Python, FastAPI, DistilBERT, XGBoost, Docker December, 2025 - February, 2026

- Reduced LLM inference costs by building a semantic routing engine that distributes queries across free-tier providers.
- **Improved inference speed by 60%** using DistilBERT-based intent classification over standard BERT
- **Achieved <2ms latency** for high-frequency queries using regex-based routing heuristics.

AstraRAG - Multi-Agent Retrieval System *Github*

FastAPI, CrewAI, LlamaIndex, ChromaDB, Docker October,2025 - November, 2025

- Designed a modular multi-agent RAG architecture separating planning, retrieval, and execution responsibilities to improve reasoning stability.
- Implemented dynamic tool orchestration using FastAPI and CrewAI for structured, controllable LLM workflows.
- Improved system reliability by ~35% through explicit agent separation and controlled retrieval pipelines.

Multi-Tenant Infrastructure as Code (IaC) Platform *Github* August, 2025- September, 2025

Python (FastAPI), Kubernetes, Docker, Helm

- **Automated provisioning of isolated environments** by developing custom Kubernetes Operators and Helm-based deployment workflows for stateful workloads.
- Built asynchronous control workflows in FastAPI to handle parallel long-running tasks efficiently.
- **Implemented containerized, scalable infrastructure** using Docker and Kubernetes to enable reliable multi-tenant workload management.

Extracurriculars

Social Media Lead, Blockchain Club,VIT

- Led promotional strategy for 4+ technical events, creating structured content plans that improved student registrations.
- Managed digital outreach generating **1,000+ cumulative impressions** across campus platforms..

Student Coordinator, Telugu Club, VIT

- Took ownership of organizing 7+ cultural events, handling everything from initial planning to final execution.
- Coordinated cross-functional student teams to manage scheduling, rehearsals, and on-ground crowd handling.

Additionals

Languages: Telugu(Native), English(Proficient), Hindi(Intermediate)