RAGHAVA N

raghavtwenty@gmail.com | +91 79049 23628 | www.raghavtwenty.com | LinkedIn | GitHub

EDUCATION

Integrated Masters in Artificial Intelligence & Machine Learning — Expected Graduation, May 2026
Coimbatore Institute of Technology, Coimbatore — *CGPA 8.7/10 as of 6th sem*

- **Related Coursework**: Mathematics and Statistics for ML, Data Structures and Algorithms, Computing in Python, Computer Networks, Operating Systems, Intelligent Agents in A, Databases and Data Management

EXPERIENCE

AI SAFETY & RESEARCH — November 2024 - Present
Anthropic — *Remote*

- Selected for Anthropic's exclusive Model Safety Private Bug Bounty Program on HackerOne, focusing on AI safety.
- Experimenting with universal jailbreaks in AI systems, specializing in CBRN (chemical, biological, radiological, and nuclear) and cybersecurity domains to exploit system refusal mechanisms.

AI RESEARCH INTERNSHIP — July 2024 - November 2024
Defence Institute of Advanced Technology — *Pune, Maharastra*

- Researched the generation and analysis of polymorphic and metamorphic malware using Large Language Models (LLMs), configured a custom malware analysis lab using FlareVM and Remnux, and performed static and dynamic malware analysis on 50+ malware samples in a sandboxed environment, delivering comprehensive reports on findings.
- Utilized AI agents and LangChain to demonstrate the effectiveness of different prompt types (Direct, Indirect, Adversarial) in generating malware with LLaMA 3.1 70B, achieving the highest success rates with adversarial prompts, generating 51 compiled and 35 executable code samples, showcasing advanced malware creation techniques.

PROJECTS

PROTECTION ONLINE — Working

- Created a Chrome extension with Generative AI and RAG to check regulatory compliance (DPDPA 2023, IT Act 2000) and to summarize E-commerce privacy policies with regional language translation, reducing reading time by 93%. Added a four-layered security approach (SSL certification, Google Safe Browsing, crowd sourced data, obfuscated JavaScript), achieving 89% accuracy in detecting malicious sites through deep learning methods.

GENZ HIRING — Working

- Engineered a resume analysis tool using LLM, LangChain, and Flask to provide tailored suggestions for aligning with career goals, streamlining the applicant experience by leveraging SerpAPI and web scraping techniques for personalized job suggestions, reducing time spent on job searches by 50%, and creating customized resumes to enhance success in specific job applications, making it an essential tool for students, professionals, and job seekers.

INTRUSION DETECTION PREDICTION — Working

- Developed and deployed a cybersecurity solution using machine learning to predict cyber attacks based on network packet data, employing preprocessing (EDA, cleaning, sampling, scaling, visualization) to reduce data redundancy by 39%. Utilized Naive Bayes, Random Forest, and XGBoost models for classification, with hyperparameter tuning and cross-validation, achieving an average accuracy of 94% with the XGBoost model, outperforming other models.

CERTIFICATIONS

- Red Teaming LLM applications, DeepLearning.AI — August 2024
- Google Cybersecurity Professional Certification, Coursera — April 2024
- Introduction to Cyber Intelligence, U.S. Department of Homeland Security — November 2023
- Docker Essentials: A Developer Introduction, IBM — October 2023

ACHIEVEMENTS

- First place in the Byte Sized Brainstrom Hackathon (The DoughVinci Code) hosted by Cookr — April 2024
- Secured 15th position out of 870+ teams in the National-level Dark Patterns Hackathon at IIT Varanasi — January 2024
- Achieved second place in Googleathon 2.0 organized by the Google Developers Student Club, SNS — November 2023

TECHNICAL SKILLS

Languages	Python, SQL, C, HTML, CSS, JavaScript, Java, R, React JS, MongoDB
Frameworks	Pandas, Scikit learn, Keras, LangChain, CrewAI, Flask, FastAPI, TKinter
Tools	Git Version Control, Virtual Machines, Docker, Figma, Cisco Packet Tracer
Machine Learning Concepts	Supervised and Ensemble Learning, Artificial Neural Networks, Deep Learning