K Harshit

+91-9845198405 | kappalaharshith@gmail.com | github - karash10 | LinkedIn- K Harshit | Portfolio

Summary

Computer Science undergrad at PES University specializing in **LLM security and adversarial ML**, with hands-on experience building secure backend systems and AI-driven threat detection pipelines. Skilled in designing scalable, high-performance architectures across security-focused AI systems and production-grade backend applications.

Education

PES University Karnataka, India

B.Tech in Computer Science and Engineering, CGPA: 8.54 *2023 – 2027*

Experience

Summer Research Intern – CCNCS, PES University June 2025 – July 2025

Karnataka, India

- Designed **XJailGuard**, a modular LLM security framework to detect multi-turn and cross-lingual prompt injection attacks in transformer-based models.
- Built multilingual classification pipelines using **PyTorch and Hugging Face**, incorporating sliding-window contextual memory for adversarial intent detection.
- Improved jailbreak detection accuracy from **8–12% baseline to 91–93%** through layered input/output filtering and zero-trust validation.
- Integrated **SHAP-based Explainable AI** and evaluated performance using false positive/negative rates and latency benchmarks; authored for ICICC 2025 publication.

Projects

XJailGuard: Explainable LLM Security Framework | *Python, PyTorch, NLP, SHAP*

- Architected a modular pipeline with independently configurable stages for input sanitization, intent classification, and output validation, enabling plug-and-play defense layer customization.
- Designed a cross-lingual threat detection module supporting 10+ languages by fine-tuning multilingual transformer models on adversarial prompt datasets.
- Built a SHAP-based token attribution dashboard to visualize flagged prompt regions, reducing manual audit time and improving model interpretability for security analysts.

Intelligent Research Analysis System | *Python, FastAPI, SentenceTransformers, NetworkX*

- Engineered an automated ingestion pipeline to parse and semantically embed academic research papers using **SentenceTransformers**, enabling dense vector representations for downstream analysis.
- Constructed a **directed knowledge graph** using NetworkX to map citation relationships, surface contradictions between papers, and identify unexplored research gaps.
- Exposed graph traversal and embedding retrieval via **FastAPI** endpoints, enabling structured querying of research landscapes through a clean RESTful interface.

SecureLogger: Adversarial Log Generation System | *Python, PyTorch, GANs, Flask*

- Designed a **GAN architecture** with a Generator trained to synthesize realistic server access logs that statistically mirror real traffic distributions for **cyber deception** use cases.
- Trained Discriminator on real server log datasets to enforce authenticity, iteratively refining Generator outputs to obfuscate **attacker behavioral fingerprints**.
- Deployed synthetic log generation as a **Flask microservice**, enabling seamless integration with simulated security monitoring and honeypot environments.

CTI-RAG: Cyber Threat Intelligence Retrieval System | *Python, LangChain, ChromaDB, Streamlit*

- Built a **RAG pipeline** over **CVE and MITRE ATT&CK** datasets, implementing custom document chunking strategies and vector indexing in ChromaDB for high-precision semantic retrieval.
- Designed a **LangChain**-based query engine that grounds LLM responses in retrieved threat intelligence, reducing **hallucination** and ensuring cited, verifiable outputs.
- Developed an interactive Streamlit dashboard for analysts to explore threat data, trace **attack patterns**, and surface relevant CVEs through **natural language queries**.

EventSphere: Scalable Event Booking System | *Java, Spring Boot, REST APIs, JWT, MongoDB*

- Architected a **Spring Boot** backend for event discovery, seat reservation, and ticket booking with a RESTful API layer designed for **high concurrency** and clean separation of concerns.
- Engineered **atomic seat-locking** and transactional booking workflows using **optimistic concurrency control** to eliminate race conditions and double-booking under load.
- Implemented **JWT-based authentication** with **role-based access control** differentiating permissions across users, organizers, and administrators.
- Optimized **MongoDB** schema design and API response times through strategic **indexing**, query optimization, and request validation middleware.

Technical Skills

Programming Languages: Python, Java, C, JavaScript
Machine Learning & AI: PyTorch, Hugging Face Transformers, NLP, Adversarial ML, Explainable AI (SHAP), GANs, SentenceTransformers, Vector Embeddings
Security: LLM Security, OWASP Top 10 (LLM Applications), Threat Modeling, Penetration Testing (Recon, Exploitation, Post-Exploitation), Nmap, Socket Programming
Backend & Systems: Spring Boot, FastAPI, Flask, REST APIs, JWT Authentication, MongoDB
Tools & Platforms: Git, Linux, bash, ChromaDB, LangChain, Streamlit

Achievements

Prof. C N R Rao Merit Scholarship

Certifications

Cybersecurity – Basics of Red Teaming