Vaibhav Saini
+1(623) 261-0432 | Tempe, Arizona | vaibhavbadolia@gmail.com | https://github.com/vaibhav-badoliasoft

Summary
Software and AI engineer, and a cybersecurity and cloud student, focused on building secure, scalable systems. I design backend architectures, real-time AI pipelines, and semantic retrieval systems with an emphasis on reliability and clean user experience. I actively explore machine learning deployment, cloud infrastructure, and security-first design. My goal is to build practical, innovative AI solutions that are secure and user-focused.

Education

- Master in Information Technology (Security) **May 2026 | GPA: 3.93**
  Arizona State University, Arizona, US
- Bachelors in Computer Science (Data Science) **June 2023 | GPA: 7.62**
  MMDU, Ambala, Haryana, India

Certifications

- AWS Certified Cloud Practitioner – Amazon Web Services (AWS) July, 2025
- AWS Academy Graduate – Cloud Architecting – Training Badge April, 2025

Technical Skills

- Programming Languages: Python, TypeScript, C#, Node.js, Angular, and React
- AI & ML: Retrieval-Augmented Generation (RAG), Embeddings, Vector Similarity Search, Sentence-Transformers, and Scikit-learn
- Backend & APIs: FastAPI, REST APIs, Model Deployment, Real-Time Inference
- Cloud & DevOps: AWS, Azure, and Docker
- Database Management: MySQL, MongoDB, DynamoDB and PostgreSQL
- Security & Monitoring: Network Security, Threat Detection, Logging, Observability, Wireshark, and Splunk

Projects

SecRAG: Retrieval-Augmented Generation (RAG) Backend | FastAPI, Sentence-Transformers, NumPy

- Designed and implemented a semantic document retrieval system supporting PDF ingestion, structured chunking, and dense vector indexing.
- Engineered character-based chunking with metadata tracking (char offsets, timestamps, source mapping) for traceable retrieval.
- Generated 384-dimensional normalized embeddings using MiniLM (all-MiniLM-L6-v2) for semantic similarity search.
- Implemented cosine similarity-based top-k retrieval engine using vector dot product for efficient semantic querying.
- Built modular architecture enabling scalable extension to LLM-based response generation.

PulseScore: Real-Time AI Inference & Monitoring System | FastAPI, Python, ML Deployment (In-Progress)

- Built low-latency ML inference API capable of handling real-time prediction workflows.
- Engineered structured logging and latency monitoring pipelines to track inference performance and system behavior.
- Designed modular model-serving architecture supporting scalable deployment and future model versioning.
- Implemented performance analytics pipeline to evaluate inference consistency and detect degradation patterns.
- Architected production-oriented backend emphasizing stateless design and extensibility.

Secure Investment Platform | Angular, .NET 8, Docker, Azure

- Collaborated with a team of 5 developers to plan, divide, and deliver a full stack fintech application for a client.
- Developed a fintech web application serving 100+ active users and handling 10,000+ API requests per minute.
- Containerized the system using Docker and deployed on Azure App Services, achieving 99.9% uptime and consistent scalability.
- Implemented Role-Based Access Control (RBAC) to secure transactions and prevent unauthorized actions.
- Integrated continuous monitoring and threat detection through Azure Security Center.
- Enhanced application reliability and data security through cloud-based API management and automated CI/CD deployment pipelines.
- Efficiently managed project with iterative and timely deliverables using Agile sprint methodology.