Kaushik Chaturvedula

+1 (224) 678-1562 | **kaushikchaturvedula@gmail.com** | **Portfolio** | **LinkedIn** | **GitHub** | **Leetcode**

SUMMARY

Software engineer with 1+ years of industry experience in backend and full-stack development, cloud-native systems, and AI-driven applications. Skilled in designing scalable microservices, developing RESTful APIs, and building secure, high-performance systems using AWS, GCP, and Azure. Proficient in Python, Java, C++, and JavaScript (Node.js, TypeScript, React, Next.js), with experience in Spring Boot, FastAPI, PyTorch, Docker, and distributed systems. Strong foundation in NLP, deep learning, and transformer-based architectures, with hands-on experience in LLM fine-tuning, RAG pipelines, and hybrid/semantic search using OpenSearch. Adept at developing reliable, event-driven services and real-time processing pipelines, with a focus on usability, fault-tolerance, and system optimization across diverse domains including healthcare, finance, and conversational AI.

EDUCATION

Purdue University, Indiana – Master of Science in Computer Science (Jan 2024 - May 2025 | GPA: 4.0/4.0)
National Institute of Technology Warangal, India – Bachelor of Technology in Mechanical Engineering (Aug 2018 - May 2022)
Ranked **9227 out of over 1.2 million students** (top 0.8%) in JEE Mains 2018, India's national engineering entrance exam.

EXPERIENCE

Medical Informatics Engineering (IN, U.S.) – Software Development Intern (Jan 2025 - Present)

- Built a Semantic Query Engine to enable intelligent, real-time querying of EHR and medical documents using RAG, NLP, LLMs, and hybrid search techniques.
- Leveraging transformers (DistilBERT) for query intent classification, OpenSearch for ANN-based vector search, and Redis caching for ultra-fast response times.
- Developed a secure and scalable FastAPI backend with a Next.js-based chatbot interface, allowing medical staff to interact with structured and unstructured clinical data using natural language. Optimizing retrieval pipelines with embedding models, metadata filtering, and hybrid search techniques to improve accuracy and latency.

Purdue University – Teaching Assistant and Research Assistant (Sept 2024 - Jan 2025)

- Assisted in delivering course material, grading assignments, and providing student support for Programming Language Design. Contributed to high-performance cosmological simulations research using Python on datasets >300TB in HPC environments.

Wibmo (Bengaluru, India) – Associate Software Engineer (Jul 2022 - Apr 2023)

- Developed and optimized backend services for a Risk-based Authentication Engine in a leading FinTech firm, enhancing fraud detection and prevention, including identifying money laundering and BIN attacks, ensuring secure transaction processing.
- Architected scalable microservices for high-traffic apps, improving reliability, performance, and optimizing transaction processing time.
- Developed asynchronous, event-driven systems with RabbitMQ and Kafka, achieving a 20% reduction in latency.

Freecharge (Mumbai, India) – Full Stack Developer Intern (May 2021 - Jul 2021)
- Optimized microservices, led memory caching initiatives, and improved load balancing to enhance efficiency and system reliability.

Ravgins (India) – Front-end Developer Intern (Jun 2020 - Aug 2020)
- Built web/mobile applications from scratch using various front-end frameworks (like Angular) and tools, enhancing user engagement.

PROJECTS
- **InsightAI**: Developing a Transformer-based Visual Question Answering (VQA) system combining Vision Transformers (ViT) and Large Language Models (BERT, GPT, BioBERT) for accurate answers on general images. Integrating multi-hop reasoning, retrieval-augmented generation (RAG), and hybrid search via OpenSearch to improve contextual accuracy. Built for real-time inference with FastAPI, and optimized deployment using Docker and ONNX for cloud (AWS, GCP) and edge devices.
- **Medical Chatbot & Semantic Query Engine**: Built an AI-powered conversational assistant for medical professionals to query structured and unstructured EHR and medical documents using natural language. Leveraged NLP, transformer models (DistilBERT), RAG methodologies, hybrid search via OpenSearch, and Redis caching to significantly enhance retrieval speed and query relevance. Created a FastAPI backend integrated with a Next.js frontend, implementing advanced metadata filtering and embedding-based retrieval mechanisms to achieve low latency and high accuracy.
- **LexiPhylax**: Engineered a multi-label hate speech classification system using PyTorch and Hugging Face Transformers with XLM-RoBERTa for contextualized embeddings. Integrated advanced features like label-wise class weighting, early stopping, mixed precision training (fp16), gradient checkpointing, and automatic sample-based checkpointing for training recovery. Tuned label-specific thresholds post-training to optimize per-category F1 scores across seven distinct hate-speech labels (e.g., race, religion, gender). Designed a batch inference pipeline for scalable deployment and built a full end-to-end training and evaluation framework with metrics tracking, threshold calibration, and checkpoint resumption.
- **SwiftNet**: Developed a high-performance C++ networking library designed for scalable, low-latency server applications using advanced features such as io_uring, kqueue, and coroutine-based task management. Optimized with request pipelining and virtual thread offloading techniques to ensure efficient CPU utilization and high-throughput request handling.

TECHNICAL SKILLS
- **AI/ML**: Deep Learning, NLP, Transformers/LLMs, PyTorch, RAG, Semantic Search (Opensearch, VectorDBs), Generative AI
- **Full-Stack Development**: Python (FastAPI, Django, Flask), Java (Spring Boot), JavaScript (TypeScript, Node.js, Express.js, React, Next.js), Angular, Cypress, Testing Frameworks
- **Databases & Operating Systems**: PostgreSQL, MySQL, MariaDB, MongoDB, Couchbase, Linux, Windows, MacOS
- **Backend & Systems Engineering**: Microservices, System Design, RESTful APIs, Backend Optimization, Distributed Systems, Asynchronous & Event-driven Architectures, Socket Programming, High Performance Computing, Multithreading/Parallel Processing
- **Cloud & DevOps**: AWS, Azure, GCP, Docker, Git, Redis, Apache Kafka, RabbitMQ, JIRA