Kaushik Chaturvedula

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

SUMMARY

Software Engineer with strong experience in AI/ML solutions, backend and full-stack systems focused on ML-driven automation, cloud-native architecture, and distributed computing. Proven ability to design and deploy scalable microservices and RESTful APIs across AWS, GCP, and Azure. Proficient in Python (FastAPI), Java (Spring Boot), C++, and JavaScript (Node.js, TypeScript, React, Next.js), and Docker for high-performance backend services. Strong understanding of modern ML workflows including transformer architectures, LLM pre-training, fine-tuning, and reinforcement learning, LLM-based function calling, model-context protocol (MCP), Agentic AI and AI agents, RAG pipelines, NLP, computer vision, deep learning, algorithms, data structures, and databases. Experienced in real-time data processing and event-driven architectures, delivering reliable, secure, and fault-tolerant systems across logistics, healthcare, and AI-powered decision-making.

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 – B.Tech in Mechanical Engineering (Aug 2018 - May 2022)
Ranked 9227 out of 1.2M+ 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 - May 2025)

- Built a RASS Engine for intelligent, real-time querying of EHR and medical documents using RAG, NLP, LLMs, and OpenSearch, with a scalable FastAPI backend and a Next.js-based chatbot for natural language interaction over structured and unstructured clinical data.
- Leveraged transformer models for query intent classification and named entity recognition (NER), and optimized hybrid/semantic search using OpenSearch ANN indices, embedding models, metadata filtering, and retrieval tuning to boost accuracy and reduce latency.
- Enhanced the RASS Engine by integrating Agentic AI with dynamic query planning and self-refining RAG loops, replacing traditional NER/intent pipelines with GPT-4o-based semantic retrieval for deeper document understanding beyond basic text lookups in document searches.
- Built an LLM proxy layer to seamlessly route requests across multiple local and cloud-based language models (OpenAI, Azure, Hugging Face), enabling dynamic model selection and integration tailored to different needs.

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 HPC cosmological simulations research on datasets over 300TB.

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

- Developed and optimized backend services for a Risk-based Authentication Engine, enhancing fraud detection (e.g., money laundering, BIN attacks) and secure transaction processing.
- Wrote database migration scripts for seamless data transfer from CouchbaseDB to MySQL/MariaDB.
- Designed scalable, asynchronous, event-driven microservices using Java (Spring Boot), Node.js, RabbitMQ, Kafka, and MySQL/CouchbaseDB, improving reliability and reducing transaction latency by 20%.

Freecharge (Mumbai, India) – Full Stack Developer Intern (May 2021 - Jul 2021)

- Developed MongoDB and Node.js services to fetch user records, optimized microservices, led memory caching improvements, and enhanced load balancing for greater system efficiency and reliability.

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

PROJECTS
- **InSightAI**: Built a real-time Visual Question Answering system using CLIP ViT-L/14, LLaMA-3 8B, and OpenSearch. Integrated GPT-4o for dynamic query planning in an agentic RAG loop with self-refining retrieval. Fine-tuned multimodal fusion for vision-language alignment over VQAv2. Served via FastAPI with GPU-optimized inference and scalable deployment.
- **Medical Chatbot and 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, 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**: Built an advanced multi-label hate speech classifier using XLM-RoBERTa with PyTorch and HF Transformers. Implemented label-wise class weighting, mixed-precision (fp16) training, gradient checkpointing, early stopping, and automatic sample-based checkpointing for scalable and resilient training. Tuned label-specific thresholds post-training to optimize per-label F1 scores across seven hate speech categories. Designed a full training, evaluation, and batch inference pipeline for production-scale deployment.
- **PlanPulse**: PlanPulse is a powerful and intuitive task management system built using Java Spring Boot and MongoDB designed to enhance team collaboration and productivity. Inspired by Jira, it provides users with robust tools to create boards, manage tasks, and collaborate efficiently.
- **SwiftNet**: Developed a high-performance C++ networking library designed for scalable, low-latency server applications using advanced features such as io_uring, kqueue, and modern C++ 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
- **Artificial Intelligence**: Deep Learning, Natural Language Processing (NLP), Transformers (all types and architectures), Large Language Models (LLMs), Generative AI, Pretraining and Fine-tuning, Parameter Efficient Fine-tuning (PEFT) - LORA/QLORA and prompt tuning with soft prompts, Prompt Engineering, Model Quantization, Reinforcement Learning with Human Feedback, Model Interpretability, Knowledge Distillation, Agentic AI, MCP, Computer Vision, CNNs - AlexNet, VGG, ResNet, DenseNet, and Vision Transformers (ViT), RNNs, LSTMs, GRUs, Machine Learning, Ensemble Learning - Bagging and Boosting, Decision Trees, Random Forest, SVM, KNN, Retrieval-Augmented Generation (RAG), PyTorch, Hugging Face, LangChain
- **Full-Stack Development**: Python (FastAPI, Django, Flask), Java (Spring Boot), JavaScript (TypeScript, Node.js, Express.js, Next.js, React, Angular), Cypress, Testing Frameworks
- **Databases and OS**: PostgreSQL, MySQL, MariaDB, MongoDB, Couchbase, Linux, Windows, MacOS
- **Backend and Systems Engineering**: Microservices, System Design, RESTful APIs, Backend Optimization, Distributed Systems, Asynchronous and Event-driven Architectures, Socket Programming, High Performance Computing, Multithreading/Parallel Processing
- **Cloud and DevOps**: AWS, Azure, GCP, Docker, Git, Redis, Apache Kafka, RabbitMQ, JIRA