Kaushik Chaturvedula

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

SUMMARY

Software and AI/ML Engineer with expertise in building intelligent, scalable systems that integrate machine learning, deep learning, and modern software engineering. Experienced in developing ML-driven automation, agentic AI systems, and retrieval-augmented architectures using LLMs, RAG pipelines, and computer vision. Strong foundation in full-stack and backend engineering, with a focus on high-performance, event-driven architectures and cloud-native deployment. Skilled in delivering robust, adaptable solutions for complex, real-world problems across diverse domains.

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 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 - May 2025)

- Designed and built a Retrieval-Augmented Semantic Search (RASS) Engine to enable real-time, intelligent querying of EHR and clinical documents through natural language, empowering healthcare professionals to interact with both structured and unstructured data.
- Developed a scalable FastAPI backend and integrated a Next.js-based chatbot UI for seamless user interaction, supported by OpenSearch, RAG pipelines, LLMs, and NLP techniques.
- Utilized transformer models for query intent classification and NER, and fine-tuned hybrid/semantic retrieval using ANN indices, embedding models, and metadata filtering to boost both accuracy and response speed.
- Integrated Agentic AI capabilities through dynamic query planning and self-refining RAG loops, replacing traditional NER/intent pipelines with GPT-powered semantic reasoning for deeper document understanding.
- Built an LLM proxy layer to dynamically route requests across OpenAI, Azure, and Hugging Face models, enabling flexible model selection and optimized performance across tasks.

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

- Assisted with Programming Language Design course and contributed to HPC cosmo-sim research on 300TB datasets.

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

- Engineered and optimized backend services for a Risk-Based Authentication Engine used in secure payment gateway systems, improving fraud detection for threats like money laundering and BIN attacks.
- Developed robust database migration scripts to ensure seamless data transition from CouchbaseDB to MySQL/MariaDB.
- Designed and implemented scalable, asynchronous microservices using Java (Spring Boot), Node.js, Kafka, RabbitMQ, and MySQL/Couchbase, resulting in a 20% reduction in transaction latency and increased system reliability.

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

- Developed backend services to support financial transaction systems for banking and payment operations, leveraging Node.js and MongoDB to optimize microservices, enhance caching, and improve load balancing for greater system efficiency and reliability.

Ravgins (India) – Front-end Developer Intern (Jun 2020 - Aug 2020)

- Designed and built responsive web and mobile interfaces for the influencer marketing platform Wobb from scratch, using modern front-end frameworks and human-computer interaction principles to enhance user engagement and experience.

PROJECTS

- **Agentic UAV Telemetry Assistant** (Volunteer Project with ArenaAI): As part of a volunteer initiative with ArenaAI, developed an agentic AI-powered system to interpret and analyze UAV telemetry logs through natural language interaction. Designed a modular, ReAct-style multi-agent architecture using LangGraph and LangChain, enabling dynamic query planning, step-wise tool execution, and context-aware responses. Integrated multi-tier agentic memory including short-term buffer memory, long-term FAISS-based vector memory, and LLM-driven entity tracking for persistent, semantically aware reasoning. Used GPT-4o for LLM inference and OpenAI embeddings for memory indexing. Built real-time streaming APIs with FastAPI and WebSockets, and also developed a custom Vue.js frontend to deliver an interactive agentic chatbot UI with token-by-token feedback. The system supports UAV log parsing (MAVLink), anomaly detection, and telemetry insight generation, showcasing how agentic workflows enable complex data exploration in structured, high-dimensional domains.
- **Medical Assistant**: Engineered an advanced Agentic RAG system for real-time natural language querying of EHR and clinical documents, featuring GPT-powered semantic reasoning, dynamic query planning, and self-refining retrieval loops. Developed a scalable FastAPI backend and a custom Next.js chatbot UI for seamless user interaction.
- **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.
- **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**: A high-performance C++ networking library designed for 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 & high-throughput request handling.

TECHNICAL SKILLS

- **Artificial Intelligence**: Deep Learning, Generative AI, Classical Machine Learning, Natural Language Processing (NLP), Computer Vision, Transformer Architectures, Large Language Models, Vision Models and Vision-Language Models, Retrieval-Augmented Generation (RAG), Causal Inference, Reinforcement Learning (RLHF), Prompt Engineering, Parameter-Efficient Fine-Tuning (PEFT), Optimization Techniques, Self-Supervised and Curriculum Learning, Meta-Learning, Agentic AI, Model Evaluation and Interpretability
- **AI Frameworks and Libraries**: PyTorch, TensorFlow, Huggingface Transformers, LangChain, Ray
- **Full-Stack Development**: Python (FastAPI, Django, Flask), Java (Spring Boot), JavaScript/TypeScript (Node.js, Express.js, Next.js, React, Angular), Cypress, PyTest, JUnit, Postman, Selenium, Mocha
- **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 Tools**: AWS, Azure, GCP, Docker, Git, Redis, Apache Kafka, RabbitMQ, JIRA