Karthik Reddy Vemireddy

karthikmasters444@gmail.com | +1 (945) 274 8248
karthikvemireddy18
arthik444

Education

Florida State University **Aug 2023 – May 2025**
Master of Science in Computer Science *Tallahassee, FL*
Coursework: Machine Learning, Artificial Intelligence, Advanced Databases, Software Engineering, Analytical Methods

International Institute of Information Technology **Aug 2018 – Jul 2022**
Bachelor of Engineering in Computer Engineering *Bhubaneswar, India*

Technical Skills

Languages: Python, TypeScript, JavaScript, Java, SQL, C++
Frontend: React 19, Angular, Next.js, Tailwind CSS, Redux, D3.js, HTML5, CSS3
Backend: FastAPI, Node.js, Express.js, REST APIs, Microservices, MongoDB, PostgreSQL, Firebase
AI & Machine Learning: LLMs, RAG, LangChain, Amazon Bedrock, Google Gemini, Vector DBs (Pinecone/HNSW), OCR
DevOps & Cloud: AWS (Lambda, S3, Bedrock), Google Cloud Platform (GCP), Docker, Git, Jira

Work Experience

AI Research Engineer **Aug 2023 – Present**
Florida State University *Tallahassee, FL*

- Engineered a high-throughput document processing system using **LLM agents**, reducing manual data extraction time from **hours to minutes** for thousands of healthcare documents.
- Architected a dynamic model routing system that integrates multiple LLMs, programmatically selecting the optimal model for each task to balance **cost, latency, and accuracy**.
- Built a cancer-specific semantic search engine using **HNSW graphs** and multilingual embeddings, enabling discovery of synonymic medical terms and reducing research search time by **70%**.
- Developed a scalable **ETL pipeline** using **OCR** and **GROBID** to convert raw PDFs/XMLs into structured JSON datasets, optimizing **MongoDB** indexing strategies for sub-second retrieval.

Software Development Engineer **Jan 2022 – Jul 2023**
Maximl *India*

- Engineered a "Dynamic Critical Path" feature using **Angular**, **D3.js**, and **BFS graph algorithms**, enabling project managers to visually identify bottlenecks and optimize plant shutdown timelines.
- Spearheaded the end-to-end development of a complex Shift Planning module, digitizing workforce management for large-scale industrial teams and significantly reducing manual scheduling errors.
- Delivered high-performance client interfaces across web and mobile ecosystems using **Angular**, **React**, **Ionic**, and **Android Java**, ensuring a seamless user experience for field operators.

Projects

ProCheck | *Winner - Google Cloud x Elastic AI Hackathon* | React, FastAPI, Elasticsearch, Gemini

- Developed an AI-driven clinical intelligence platform allowing customizable PDF ingestion, automatic embedding generation, and real-time streaming of actionable medical checklists.
- Engineered **Elasticsearch Hybrid Search** by combining BM25 keyword precision with vector embeddings using **Reciprocal Rank Fusion (RRF)**, significantly improving retrieval relevance for ambiguous medical queries.
- Integrated **Google Gemini** for query expansion and context-aware checklist generation, orchestrated via a **FastAPI** backend with Pydantic validation and **Firebase** for conversation persistence.

LeetSpace (myleetspace.com) | *Founder & Full Stack Engineer* | React 19, FastAPI, MongoDB, Firebase

- Launched a production-grade coding interview platform transforming ad-hoc practice into systematic learning via a custom **spaced-repetition algorithm** and active recall workflows.
- Architected a high-concurrency **FastAPI** backend with **Async I/O** and **MongoDB (Motor)**, enabling complex features like solution version control, mistake tracking, and "retry later" queues.
- Built a performance-optimized **React 19** frontend featuring an in-browser code editor (**CodeMirror**), interactive activity heatmaps, and secure token-based authentication.

Rift Rewind (AI Coach) | *Python, React, AWS Bedrock, Claude 3, Riot Games API*

- Implemented a conversational coaching agent using **AWS Bedrock (Claude 3)** with **function calling**, enabling the AI to autonomously query real-time analytics to benchmark progress and generate practice plans.
- Engineered an end-to-end data pipeline that ingests raw **Riot API** match history to compute derived features (vision trends, objective control), synthesizing them into structured insights and narratives.
- Optimized application latency using aggressive **client-side caching** and pre-aggregated analytics, allowing users to navigate full-year performance timelines instantly.