# Harish Kumar Manepalli

951-538-1306 linkedin.com/in/harish-kumar-manepalli harish.jobs.applications@gmail.com harishkumar.work

## Summary

Software Engineer with 6+ years of experience building scalable, distributed systems and cloud-native applications while driving impactful innovations in AI. Proven track record of designing low-latency, fault-tolerant architectures with **Java, Spring Boot, React, AWS, and Kafka**, serving 100,000+ users and processing 100K+ daily transactions. Adept at modernizing legacy monoliths into microservices, deploying on Kubernetes, and implementing CI/CD pipelines for reliable, production-grade delivery.

In parallel, contributed to cutting-edge AI research and applications by fine-tuning transformer-based models such as **GPT-2 Large, LLaMA 3.1 (8B), and GPT-3.5**, and applying **BERT, Word2Vec, and CLIP to real-world problems**. Built custom **CNNs and transformer architectures, optimized deep learning pipelines**, and deployed models in production with robust cloud infrastructure. Delivered measurable impact, which saved millions of dollars and thousands of man-hours by streamlining business processes and modernizing legacy systems.

Passionate about solving complex, ambiguous challenges across both large-scale software engineering and applied AI, with a focus on scalability, performance, and real-world impact.

## Experience

**Capital One (Client)** May 2025 - Present

*Software Engineer, AI* *Riverside, USA*

- Developed an **AI-driven analytics agent** using **LangChain and OpenAI APIs** to transform plain-text queries from product owners into dashboard-ready insights, cutting ad-hoc analytics turnaround time by 70%.
- Leveraged **BERT-based** embeddings for intelligent customer query routing, improving chatbot resolution accuracy by 15% and reducing call center volume.
- Built and maintained REST APIs to support customer credit workflows, enabling integration across 15+ internal platforms and handling peak loads of 20K RPS.

**Islanders** Aug 2024 - May, 2025

*Software Engineer* *Riverside, USA*

- Spearheaded the transformation of a static site into a distributed e-commerce platform using **React, Spring Boot, and MySQL**; integrated Kafka for real-time messaging, reducing transaction processing time to under 1 second
- Designed and deployed scalable inventory management and threshold-based alerting system across **AWS (S3, EC2, RDS, SNS, ASG, Route 53)** enhancing reliability and system uptime
- Architected cloud-based distributed services handling high concurrency and ensuring seamless auto-scaling and fault tolerance under load

**Fidelity Investments** Jul 2018 – Aug 2022 (4 years)

*Software Engineer | Full stack development* *Bengaluru, India*

- Gained over 4 years of experience in: writing clean and scalable code; developing front-end and back-end features for large-scale applications following CI/CD; working with project stakeholders to gather project requirements, conducting peer code reviews
- Developed scalable, high-performance, and responsive web applications using **Angular, React, React Native, TypeScript, JavaScript, and Material UI** used by 100,000+ employees. Built reusable UI components, developed cross-platform applications, implemented client-server HTTPS communication via OAuth
- Developed high-performance, low-latency **RESTful services using Java, Spring Boot** using data structures, algorithms, and concurrency mechanisms. Used automated testing frameworks like **JUnit** achieving 95% test coverage to ensure code quality, and maintainability
- Managed **relational (MySQL) and NoSQL** databases, writing PL/SQL queries, stored procedures, and managing transactions to ensure data consistency, handling over 100,000 daily transactions
- Collaborated with 10 developers to redesign an application from **Monolith to micro-services architecture** and used **Kubernetes to deploy these services to the AWS cloud** to increase reliability, scalability, and security
- Implemented a **Python Parser** to extract trades from employees' quarterly report PDFs for checking employee compliance, thereby streamlining a manual process which was estimated to save the organization a 2 million USD
- Conceptualized and built a **AI chat bot using elastic search** to assist employees in finding correct requests, decreasing average ticket volume for backend manual processing team by 15%
- Engaged in **production support** rotations, performing root cause analysis, troubleshooting distributed logs via **Splunk**, and improving service uptime

## Skills

**Programming Languages**: Java, Python, C, C++, C#, SQL, PHP, JavaScript, TypeScript, HTML, CSS, Shell Scripting
**Technologies & framework**: Spring, Spring Boot, Hibernate, Angular, React.js, React Native, Node.js, .NET, RESTful APIs, gRPC, Kafka, JUnit, Jest, SOAP
**Cloud & DevOps**: Amazon Web Services (S3, EC2, RDS, Lambda, SNS, ASG, Route 53), Kubernetes, Docker, Jenkins, Git, uDeploy, CI/CD pipelines, Terraform, Model Deployment on Cloud (SageMaker, EC2)
**Databases**: MySQL, PostgreSQL, MongoDB, Redis, Apache Cassandra
**AI/ML & Data Science**: Large Language Models (GPT-2, GPT-3.5, LLaMA 3.1, BERT, Word2Vec), LangChain, LangGraph, HuggingFace Transformers, TensorFlow, PyTorch, Scikit-Learn, OpenAI API, RAG pipelines, Multi-agent frameworks, Fine-tuning & LoRA, Model quantization (INT8, FP16), Transfer Learning, CNNs, Transformers, Topic Modeling (LDA, BERTopic)
**Data Engineering & Analytics**:Vector Databases (FAISS, Pinecone), Elasticsearch, PyLucene, Pandas, NumPy, Matplotlib, Seaborn, Plotly, Tableau, Power BI, Data Preprocessing & Feature Engineering, SQL query optimization, Data Pipelines (ETL/ELT), Statistical Modeling, Experimentation & A/B Testing

## Education

**University of California, Riverside** — Riverside, CA
*Master of Science in Computer Science; CGPA - 3.9/4.0* — *Sept. 2022 – Mar 2024*

**National Institute of Technology, Tiruchirappalli** — India
*BTech. in Instrumentation and Control Engineering* — *Aug. 2014 – May 2018*

## Projects

**Privacy Preserving Machine Learning (PPML) - CRYPTGPU (Open-Souce)** Sept 2023 - Mar 2024

- Worked on a novel project aimed at speeding up machine learning training on encrypted data by shifting all the operations to GPU. This has accelerated training by 8 times and will help in scalability of privacy-preserving ML models, thereby making ML solutions more secure

**Topic Modeling (Novel Project)** Apr 2023 - Jun 2023

- Analyzed different topic modeling models for research paper abstracts using the prominent methods: TF-IDF, BERT, LDA, BERT+LDA. The BERTopic model performed well in terms of coherence (0.65) and silhouette score (0.52), indicating strong semantic similarity among words in the topics, and well-separated topic clusters.

**Web Search Engine (similar to Google)** Jan 2023 - Mar 2023

- Scraped 500MB sports articles. Used PyLucene to remove stop words and create an optimised index, and searched on this index with user query for relevant pages. Leveraged BERT to generate embeddings of the data, and the query embeddings. They were compared using cosine similarity to retrieve the most relevant data.

**Hazard Free Navigation (similar to Google Maps)** Apr 2023 - Jun 2023

- Developed a web application that avoids crime hotspots and directs users through a safer path in SF City. Generated 1000 hotspots from the crime dataset using DB Clustering. Employed OSMnx to extract street network from OpenStreetMap and removed hotspot nodes, shortest path was calculated with A* Heuristic Algorithm

**LLM Fine-Tuning and Deployment for Conversational AI** Apr 2023 - Jun 2023

- Fine-tuned gpt2-large and LLaMA 3.1 (8B) models using domain-specific datasets to build a lightweight, low-latency chatbot for enterprise internal tools.Applied low-rank adaptation (LoRA) and quantization (INT8) techniques to reduce inference cost by 50%

**CalcGPT: Arithmetic with GPT-2 Large** Sep 2023 - Dec 2023

- Fine-tuned GPT-2 Large to perform basic arithmetic operations, evaluating its capability to learn mathematical patterns without explicit calculators. Explored the model's performance using creative prompts, analyzing its accuracy and limitations in handling numerical computations

## Other Experience

- Applied Large Language Models (LLMs), deep learning frameworks and libraries (PyTorch, TensorFlow) to develop ML models. Implemented transfer learning strategies, effective regularizers and Data augmentation strategies and attained state-of-the-art model accuracies of over 95% for a range of NLP and Computer Vision tasks