YASHWANTH REDDY DASARI

312-646-8975 | yashwanth2632@gmail.com | Portfolio | LinkedIn | Github

Full-Stack Software Engineer skilled in building scalable distributed systems and AI-powered applications from 0-to-1. Proficient in backend architecture (Java, Python, Go, Node.js) and modern frontends (React, TypeScript). Proven ability to work directly with customers, own product decisions, and ship high-impact features in fast-paced startup environments.

EDUCATION

University of Illinois Chicago (*Master of Science in Computer Science) CGPA - 3.9/4* — Aug 2023 – May 2025

Vellore Institute of Technology (VIT) (*Bachelor of Science in Computer Science) CGPA - 3.7/4* — July 2019 – May 2023

WORK EXPERIENCE

ImpacterAI Inc, *Software Engineer, San Francisco, CA* — Aug 2025 – Present

- Architected a multi-agent orchestration engine for a real-time conversational AI platform using Python, FastAPI, and async I/O with Redis pub/sub, processing 5M+ daily calls with sub-second response times across a distributed, fault-tolerant environment.
- Collaborated with ML teams and worked directly with enterprise customers to build an end-to-end inference pipeline (OpenAI GPT-4, FastAPI, React/WebSockets), shipping features from prototype to production in 6 weeks based on customer feedback.
- Optimized high-throughput ML serving infrastructure on Azure (AKS, Azure ML) by implementing KV caching and request batching, reducing memory consumption by 68%, improving p99 inference latency by 70%, and cutting inference costs 40%.
- Took ownership of the rapid prototyping cycle, working directly with the founder to ship 3 production features in the first month, establishing API design patterns adopted across the engineering team.
- Engineered a full-stack observability platform, using Go for high-throughput event ingestion (80K+/min) and a Node.js API powering a real-time React/TypeScript dashboard, reducing incident detection time by 60% across 12 microservices.

University of Illinois at Chicago, *Software Engineer, Chicago, IL* — Nov 2023 – Aug 2025

- Owned product direction for a semantic search platform, conducting user research with 50+ researchers to identify pain points, driving 90% adoption & securing its use as the primary research tool.
- Led the 0-to-1 technical implementation using React, TypeScript, Node.js/Express, Elasticsearch, and a LangChain RAG pipeline with vector similarity, enabling sub-second retrieval across 62M+ records & accelerating research timelines by 10x.
- Designed RESTful & gRPC microservices using Java Spring Boot/FastAPI with circuit breakers, handling 6M+ daily requests with sub-second latency and 99.95% uptime across 8 containerized services using multi-layer caching (Redis, MongoDB).
- Optimized HPC workflows on AWS (24+ A100s) for AI training using PyTorch DDP by rewriting CUDA kernels & enabling model parallelization, cutting training time 67% (6→2 hrs).

SimplyFI Innovations, *Software Engineer 1, India* — May 2022 – Aug 2023

- Partnered with finance teams to architect & deliver an intelligent document processing solution (Python, FastAPI, GPT-3.5, Hugging Face) with 94% accuracy, reducing manual review time by 90% (2 days → 3 hours) & accelerating client approvals.
- Built a real-time fraud monitoring dashboard (React, Node.js, WebSockets, Prometheus) processing 10M+ daily transactions, cutting incident detection time by 67% & minimizing financial loss.
- Engineered a CI/CD automation system on AWS (ECS, ECR) with Docker, Kubernetes, and Terraform, reducing deployment cycles from 2 days to 1 hour across 16 microservices, improving developer velocity & enabling zero-downtime deployments.

PROJECTS

Real-Time AI Bias Mitigation (LLM-as-Judge) | **Paper**

- Built Bias Mitigation & AI safety framework using LLM-as-Judge (DeepSeek) & an inference-time ”Fairness Mediator”.
- Developed a React/WebRTC playground demonstrating the ”Mediator” correcting bias via Adversarial Debiasing in real-time.
- Achieved +66.7% Bias Safety Score on Mistral-7B, 0.87 human-eval (Cohen’s K), & halved 7 types of jailbreak attacks.

LLM Optimization (Instant Soup Pruning + QLoRA) | **Paper**

- Architected LLM Pruning (ISP) + Quantization (QLORA) pipeline; cut parameters 30% & improved perplexity 65%.
- Built React/FastAPI demo UI visualizing 32% inference speed-up & 55% memory reduction on Gemma-3-1B & Mistrail.
- Reduced PEFT training time 77% & cut energy use 32% for sustainable, cost-efficient model deployment.

TECHNICAL SKILLS

- **Languages:** Java, Python, Go, C#, C++, JavaScript, TypeScript, SQL, Bash, CUDA.
- **AI/ML:** RAG, LangChain, OpenAI API, Prompt Engineering, Pruning, Quantization, PyTorch, TensorFlow, Hugging Face.
- **Backend:** FastAPI, Spring Boot, Node.js/Express, Microservices, RESTful APIs, gRPC, Hibernate, GraphQL, Django.
- **Frontend:** ReactJS, Next.js, AdonisJS, Redux, TypeScript, WebSockets, WebRTC, D3.js, Server-side events.
- **Databases:** PostgreSQL, MongoDB, Redis, MySQL, Elasticsearch, Vector Databases, DynamoDB
- **DevOps & Cloud:** AWS, Azure, GCP, Docker, Kubernetes (Production), CI/CD, Terraform, GitHub Actions, NGINX, Linux.
- **Tools:** Kafka, RabbitMQ, Prometheus, Grafana, Spark, Git, npm, Postman.
- **Certifications :** AWS Solutions Architect – Associate (2024), NVIDIA-Certified Associate: Multimodal Generative AI