Harsha Vardhan Yellela

United States | +1-248-497-9965 | harshavardhanyellela7@gmail.com | har5ha.in | LinkedIn | GitHub

Summary

Full Stack Engineer with 3+ years of experience building production-grade, cloud-native systems across Node.js, Python (FastAPI), Next.js, and Java/Spring Boot. Currently building voice and SMS agentic agents at teli.ai using OpenAI GPT-4o function calling, Twilio, and LangChain — deployed for mortgage clients including bevri.ai and NEXA Lending. Strong background in microservices, REST/GraphQL APIs, RAG pipelines, multi-tenant SaaS architecture, and AWS-native deployment with Docker, Kubernetes, and CI/CD automation.

Skills

- **Languages:** Python, TypeScript, JavaScript, Java, Go, SQL
- **Backend:** Node.js, Express, FastAPI, Spring Boot, REST/GraphQL APIs, Microservices, OAuth2/JWT
- **Frontend:** Next.js, React, TypeScript, Tailwind, Component-Based Architecture, Responsive UI
- **AI/ML & LLM:** OpenAI GPT-4/4o Function Calling, LangChain, LangGraph, CrewAI, RAG, AWS Bedrock, pgvector
- **Voice & Telephony:** Twilio Voice + SMS, 10DLC Compliance, Real-Time Transcription, Speech-to-Text
- **Cloud & DevOps:** AWS (EKS, ECS, Lambda, S3, RDS, Bedrock, API Gateway), GCP, Docker, Kubernetes, Terraform
- **CI/CD:** GitHub Actions, Jenkins, Multibranch Pipelines, Automated Build/Test/Deploy, IaC
- **Databases:** PostgreSQL, Supabase, DynamoDB, MongoDB, Redis, pgvector, Query Optimization
- **Event & Messaging:** Apache Kafka, Event-Driven Architecture, SMTP, Webhooks
- **System Design:** Multi-Tenant SaaS, High Availability, Fault Tolerance, Distributed Systems, Performance Tuning
- **Observability:** Prometheus, Grafana, ELK Stack, Logging, Metrics, Alerting, MTTR Reduction
- **Practices:** Agile/Scrum, TDD, Unit & Integration Testing, Code Reviews, CI/CD Automation

Experience

Full Stack Engineer – teli.ai, USA **Apr 2026 – Present**

- Built voice calling agents with OpenAI GPT-4o function calling and Twilio telephony, to automate inbound/outbound mortgage loan officer calls with real-time transcription and summarization, to auto-qualify leads and surface qualified prospects back to officers, for clients including bevri.ai and NEXA Lending in production.
- Engineered SMS agentic workflows with 10DLC compliance setup, agent-creation pipeline, and conversation state management in Node.js, to enable automated two-way lead nurturing at scale, to remove manual rep effort from outreach, for mortgage broker clients live in production today.
- Integrated teli.ai voice and SMS agentic capabilities into bevri.ai via REST APIs and Next.js + Node.js services, to deliver a unified AI-native CRM for loan officers, to extend teli.ai's product surface into the mortgage CRM vertical, for bevri.ai's imminent customer launch.
- Implemented multi-tenant data isolation in bevri.ai using row-level security on Supabase/PostgreSQL, to enable per-loan-officer and per-brokerage workable environments, to safeguard mortgage PII and meet compliance requirements, for the bevri.ai CRM platform.
- Built RAG pipeline ingesting chat logs and call transcripts from teli.ai into pgvector with LangChain orchestration, to expose retrieval-augmented Q&A over historical context, to give loan officers instant access to past borrower interactions, for the bevri.ai agent stack.
- Implemented SMTP integration with bring-your-own-domain support, to route inbound mortgage leads directly to loan officer inboxes through their own domains, to remove platform-as-middleman friction, for white-labeled mortgage workflows on bevri.ai.
- Owned end-to-end deployment on AWS (ECS, EKS, Lambda) with Dockerized services, Kubernetes manifests, and GitHub Actions/Jenkins CI/CD pipelines, to support daily and weekly releases at 99%+ uptime, to keep the engineering org shipping fast, for production teli.ai infrastructure.

Software Engineer Research Assistant – Agentic AI | Lawrence Technological University, USA **Jan 2025 – Dec 2025**

- Designed multi-agent orchestration systems using CrewAI and LangChain, to automate research workflows like literature review and data extraction, to reduce manual effort by 70%, for university research teams.
- Built and deployed 3 persistent AI agent services on AWS EKS with Fargate, to enable scalable execution of agent pipelines, to support semantic search across 10K+ documents with sub-second latency, for academic research applications.
- Developed FastAPI and GraphQL backend services integrated with AWS Bedrock, to expose LLM-powered capabilities, to deliver dynamic query responses and document summaries, for AI-driven research platforms.
- Implemented OpenSearch Serverless and Redis caching layers, to optimize document retrieval and session management, to enhance system responsiveness and reduce redundant LLM calls, for high-frequency research queries.
- Automated infrastructure provisioning using Terraform and GitHub Actions CI/CD, to streamline environment setup, to reduce provisioning time from 2 hours to 15 minutes, for cloud-native deployments.
- Conducted benchmarking between no-code (n8n) and coded (CrewAI) agent frameworks, to evaluate performance trade-offs, to guide architectural decisions for follow-on AI research initiatives, for faculty and engineering stakeholders.
- Built scalable backend pipelines in Python with asynchronous processing, to handle concurrent agent tasks, to improve execution efficiency and system throughput, for multi-agent AI workloads.

SDE-1 (LN Technical Consultant) | Infor India Pvt. Ltd., Hyderabad, India **Apr 2022 – Dec 2023**

- Developed enterprise-grade RESTful APIs using Java and Spring Boot, to integrate ERP systems with external clients (Ferrari, Boeing, Triumph), to process 500+ daily transactions reliably, for global manufacturing and supply chain systems.
- Engineered serverless microservices using AWS Lambda, API Gateway, and S3 event triggers, to automate data ingestion and asynchronous event processing, to improve scalability and reduce infrastructure overhead, for enterprise clients.
- Diagnosed and resolved 15+ critical data pipeline issues across MySQL and distributed ERP systems, to stabilize batch processing workflows, to reduce failure frequency from weekly to monthly, for production operations teams.
- Built backend services in C++ for performance-critical modules, to optimize data processing and reduce latency in ERP integrations, to enhance system efficiency for high-volume enterprise workloads.
- Containerized applications using Docker and implemented CI/CD pipelines, to standardize deployments across client environments, to reduce release turnaround time by 25%, for multi-client delivery teams.
- Designed frontend components using React.js and JavaScript, to improve user interaction with ERP dashboards, to enhance usability and reporting efficiency, for business users.
- Collaborated in Agile teams to deliver features across microservices and distributed architectures, to ensure timely delivery and system reliability, for enterprise stakeholders and client integrations.

Student Consultant – AI/ML Internship | Open Avenues (Build Fellowship), Remote **Feb 2026 – Mar 2026**

- Built a 10.81M-parameter GPT decoder-only transformer from scratch in PyTorch with multi-head causal self-attention, residual connections, and pre-LayerNorm, to validate end-to-end transformer training on dual RTX 3090 GPUs, to reach val loss ~1.18 at 25K steps generating coherent dialogue, for the LuffyGPT fellowship deliverable.
- Trained a custom BPE tokenizer via SentencePiece on a ~3.2M character corpus (vocab 2000, 3.11× compression), to benchmark char-level, GPT-2, cl100k, and o200k tokenization strategies, to inform tokenizer choice for downstream training runs, for the fellowship's project evaluation.
- Deployed the trained model as a Gradio app on HuggingFace Spaces with weights hosted on HuggingFace Hub and a CLI supporting train/eval/interactive modes, to enable public interactive inference and reproducible runs, to demonstrate end-to-end ML delivery, for fellowship review and portfolio presentation.

Projects

Resume Optimizer – QLoRA Fine-tuned LLM for ATS Optimization **GitHub**

PyTorch, QLoRA, PEFT, TRL, Transformers, FastAPI, Ollama

- Fine-tuned Qwen3-4B with QLoRA (4-bit NF4, LoRA rank 16, alpha 32) on 1,304 examples curated from 1,800+ resumes, to specialize the model for ATS-aware resume generation, to achieve 9.5/10 quality score on GPT evaluation at 18–22GB peak VRAM, for job-seekers needing role-tailored resumes.
- Built FastAPI inference service with structured JSON output and 3–5s response time on RTX 3090, to expose the fine-tuned model behind a REST API, to enable automated resume generation pipelines, for downstream applicant tracking workflows.

ML Sentiment Feedback Loop – Production MLOps Platform **GitHub**

AWS (ECS Fargate, SageMaker, S3), Terraform, GitHub Actions, Docker

- Architected 8-microservice MLOps platform with an automated Inference → Feedback → Evaluation → Retraining → Deployment loop, to remove manual intervention from model lifecycle, to enable continuous improvement from production signal, for ML teams running customer-facing sentiment workloads.
- Configured SageMaker auto-retraining triggers, model registry versioning, SNS deploy notifications, and Terraform IaC with GitHub Actions CI/CD to ECS Fargate, to ship infra and code from a single PR, to enable safe, repeatable rollouts, for production MLOps environments.

Lambda Microservices Platform – Enterprise Serverless Backend **GitHub**

Python, AWS Lambda, DynamoDB, API Gateway, React, Next.js, Terraform

- Architected 94 AWS Lambda functions forming a complete SaaS backend with REST and GraphQL APIs via API Gateway, to support a field-service product end-to-end, to scale on demand without standing infrastructure, for enterprise customers in production.
- Integrated 10+ third-party services (Stripe, Twilio, DocuSign, QuickBooks) with a React/Next.js admin dashboard and React Native mobile app (104+ components), to deliver a unified operator experience, to consolidate billing, signing, and comms in one platform, for field-service operators and admins.

Education

Master of Science in Computer Science · GPA: 3.77/4.0 **Jan 2024 – Dec 2025**

Lawrence Technological University, Southfield, MI

- **Relevant Coursework:** Machine Learning, Artificial Intelligence, Natural Language Processing, Intelligent Robotics (ROS), Collaborative Research in Agentic AI