# AYUSH SOLANKI

**Machine Learning Engineer**

(+91) 701-676-9052 ayush17solanki@gmail.com LinkedIn: ayushsolanki17 GitHub: AyushSolanki-17
Portfolio: ayushsolanki.dev Location: Ahmedabad, Gujarat, India, Open to remote work

## Experience

**Optimoz Engineering - AI Driven Healthcare Solutions** *AI/ML Engineer* **Sep 2024 – Present**

- Fine-tuned and evaluated **LLaMA and DeepSeek** models for domain-specific healthcare Q&A using curated test cases; improved answer accuracy by **26%** on held-out medical query benchmarks and decreased response inconsistency by **30%** through systematic prompt engineering and output calibration.
- Designed and implemented a **unified LLM tool-calling gateway** supporting **4 providers** (OpenAI, Gemini, Grok, DeepSeek) with standardized function schemas, dynamic routing, and automatic fallbacks; reduced new model integration time from **2 days to < 2-4 hours** through abstracted interface patterns.
- Architected production-grade **Retrieval-Augmented Generation (RAG)** pipelines using hybrid search (BM25 + dense embeddings) with cross-encoder re-ranking; achieved **35–40% hallucination reduction** and **25% improvement in retrieval precision@5** on internal knowledge base queries through iterative chunk optimization.
- Built **agent-based execution workflows** enabling automated code generation and multi-step tool orchestration across LLM providers; lowered manual operational effort by providing JSON based schema for workflows. Developed a scalable **workflow orchestration framework** supporting multi-agent directed acyclic graphs, conditional routing, and user-scoped context isolation with MCP integration.

**KenexAI** *AI/ML Engineering Intern* **Jan 2024 – Sep 2024**

- Developed an end-to-end **NLP-to-SQL system** enabling natural language queries over relational databases; achieved **85% query correctness** on production schemas with **complex multi-table joins** through schema-aware prompting, query validation, and iterative refinement.
- Built an **LLM-powered meeting summarization system** generating structured outputs (summaries, action items, technical notes) from zoom calls and transcripts; reduced post-meeting documentation time by **70%** through template-based generation and multi-stage extraction pipelines.

## Education

**Government Engineering College, Gandhinagar** **2021 – 2024**
Bachelor of Engineering in Information Technology **CGPA: 9.07**

## Technical Skills

*Languages:* Python, C++, Go |*Backend:* FastAPI, Flask, REST APIs, Authentication
*ML & LLMs:* Transformers, Large Language Models, Fine-Tuning (LoRA), RAG pipelines, embeddings
*Infrastructure:* AWS, Docker, Kubernetes, CI/CD pipelines, containerized deployments, service orchestration
*Databases & Search:* PostgreSQL, Weaviate, BM25, HNSW, vector search, hybrid retrieval systems
*Frameworks:* PyTorch, Tensorflow, LangChain, LangGraph, experiment tracking, model integration
*Systems:* Distributed systems, concurrency, deterministic execution, fault tolerance, caching, consistency models

## Projects

**MiniVec: High-Performance HNSW Vector Search Engine** GitHub

- Implemented a from-scratch **HNSW index in C++17** with multi-layer graph construction; achieved **89% recall@10** on a **1M-vector benchmark** with **11.7ms P50 latency**, outperforming brute-force search by **2.6x**.
- Designed a thread-safe, deterministic build pipeline and exposed zero-copy Python bindings via pybind11, enabling seamless integration into ML pipelines with predictable latency and reproducible builds.

**StepEngine: Distributed Serverless Workflow Orchestration Engine** GitHub

- Built a **distributed workflow engine in Go** supporting DAG execution, retries, fan-in/fan-out, and conditional routing with **exactly-once execution semantics** using PostgreSQL transactional state transitions.
- Developed a **deterministic scheduler** with O(N + E) complexity, reducing database load and reconciliation overhead for large workflows.

## Research and Publications

Class-Conditional Regularization for Cross-Lingual Representation Stability (Under Review at PRL)
Enhancing Emotion Recognition Using Multimodal Deep Neural Networks [DOI]
Efficient Join Operations for Utility List-Based High-Utility Mining [DOI]

## Certifications

AWS Certified Solutions Architect [Certificate] Google Solution Challenge – Global Top 100 [Certificate] UNESCO Hackathon Finalist [Certificate] Smart India Hackathon Winner [Certificate]