

NIKHIL SHINGADIYA

Ahmedabad, Gujarat, India | +91 6354071470 | shingadiyanikhil1634@gmail.com | [LinkedIn](#) | [GitHub](#)

ABOUT ME ([Portfolio](#))

Machine Learning Engineer with 4-5 years of experience building and deploying production-grade AI/ML systems, specializing in Generative AI, Large Language Models, RAG, and multi-agent architectures. Experienced in designing autonomous AI workflows, LLM routing and failover systems, model serving with vLLM, secure code-execution infrastructure, and scalable AI backends. Strong background in Python, FastAPI, distributed workers, vector databases, cloud infrastructure, and LLM fine-tuning, with a track record of improving automation, reliability, performance, and operational efficiency.

TECHNOLOGY

Generative AI & LLMs: Generative AI, LLMs, Transformers, RAG, Prompt Engineering, LLM Fine-Tuning, Tool-Augmented LLMs, Agentic AI, Multi-Agent Systems, NLP.

AI/ML: Machine Learning, Deep Learning, Time-Series Forecasting, Classification, Regression, Clustering, Statistics, Probability, Hypothesis Testing, Bayesian Analysis, PyTorch, TensorFlow, Scikit-learn.

LLM & AI Frameworks: Hugging Face, LangChain, LangGraph, LlamaIndex, vLLM, Ray.

Backend & Distributed Systems: Python, FastAPI, Celery, ARQ, Redis, AsyncIO, REST APIs, Docker, Docker Compose

Data & Retrieval: Pandas, NumPy, PySpark, MongoDB, PostgreSQL, MySQL, Milvus, Pinecone, Qdrant, Vector Search, Data Pipelines.

Cloud & Infrastructure: AWS EC2, S3, ECR, Lambda, CloudWatch, Linux, Git, GitHub, CI/CD

Data & Visualization: SQL, Web Scraping, BeautifulSoup, Selenium, Matplotlib, Seaborn, Plotly

Development: OOP, System Design, Microservices, Distributed Processing, VS Code, PyCharm, Jupyter

WORK EXPERIENCE

Rishabh Software — Ahmedabad, Gujarat

Machine Learning Engineer | July 2025 – Present

- **GO AI Project:**
- **Objective:** Engineered a FastAPI-based AI backend supporting a Node.js/Electron desktop browser application, providing production-grade APIs for multi-agent orchestration, autonomous output generation, validation, and execution workflows.
- **Tech Stack:** FastAPI, Qwen 30B, vLLM, Docker, MongoDB, Redis, ARQ Workers, AsyncIO, Pandas, NumPy, HTML/CSS, JS, Mermaid, and various REST APIs (FRED, World Bank).
- **Key Achievements:**
- Architected a production **multi-agent AI system** with 5+ specialized agents for autonomous planning, execution, validation, and output generation, reducing task completion time by 45%.
- Engineered intelligent **LLM routing** and failover mechanisms across multiple models, achieving 98% service availability and resilient recovery from model failures and latency spikes.
- Built **Data George**, an AI-driven data orchestration layer that discovers, retrieves, and normalizes data from 20+ external sources including FRED, World Bank, and arXiv, supplying validated datasets to downstream Python execution jobs.
- Developed a secure sandboxed **code-execution** microservice using FastAPI, Redis, ARQ, and Docker, processing 10,000+ daily Python jobs with isolated execution, automated error detection, and self-repair workflows.
- Built a **multi-surface generation** engine that transforms agent outputs into interactive HTML layouts, visualizations, diagrams, and technical documentation, reducing manual content-generation effort by 70%.
- Skilled in constructing Large Language Models (**LLMs**) and fine-tuning them for precise applications, specifically in generating stories and summarizing complex narratives.

- **BI-Agent:**
- **Objective:** Designed and developed a multi-user **AI-powered Business Intelligence** platform that automatically extracts database schemas and metadata, converting natural-language questions into SQL queries, visualizations, and analytical insights with role-based access control (**RBAC**).
- **Tech Stack:** FastAPI, PostgreSQL, SQLAlchemy, Celery, Redis, LangGraph, Milvus, MinIO, OpenAI, and Docker. Intel OPEA Architecture.
- **Key Achievements:**
- Developed an LLM-powered natural-language-to-SQL pipeline **achieving 85%** query-generation accuracy, enabling users to interact with enterprise databases using natural language.
- Applied the **Elbow Method** on retrieval scores to optimize similarity thresholds, cutting token costs by 30% and reducing noise in document searches.
- Integrated Intel **OPEA components** to automate database schema extraction and metadata generation, creating reusable pipelines for enterprise data discovery.
- Built a LangGraph-based **FAQ generation** workflow that analyzes database relationships and generates complexity-controlled contextual examples for improved NL-to-SQL understanding.
- Engineered a modular FastAPI backend with Celery and Redis for asynchronous AI processing, schema analysis, and **metadata generation** workflows.
- Implemented Milvus for vector-based metadata retrieval and MinIO for **self-hosted object storage**, supporting privacy-sensitive enterprise deployments.
- Developed an LLM-powered visualization **ECharts** engine that identifies SQL query results and generates interactive data visualizations for business analysis.
- Deployed on self-hosted Rishabh servers, ensuring data privacy, scalability, and complete operational control.

Unlimited WP — Ahmedabad, Gujarat

AI/ML Lead Developer | Dec 2023 – July 2025

- **WeamAI:**
- **Objective:** Engineered an AI-driven ecosystem for enterprise onboarding and collaborative document intelligence, leveraging advanced LLMs to optimize organizational workflows.
- **Tech Stack:** FastAPI, Celery, Redis, Docker Compose, LangChain, LlamaIndex, Ray, Pinecone, Qdrant, OpenAI, HuggingFace, Anthropic, Streamlit, and AWS EC2.
- **Key Achievements:**
- **Orchestrated multi-container** deployment using Docker, ensuring high availability and seamless scalability across production environments.
- Designed a modular Layered Framework to decouple core services, significantly improving **system maintainability** and long-term codebase health.
- Developed automated **CI/CD workflows** for rapid iteration, integrating proactive monitoring to slash release cycles and improve reliability.
- Built a specialized **Retrieval-Augmented** Generation system, achieving superior factual grounding and context-aware responses for enterprise data.
- Deployed LangChain agents to automate complex document navigation and provide interactive, multi-step user guidance.
- Employed Pinecone and Qdrant for efficient storage and retrieval of embedding vectors, optimizing AI model performance.
- Utilized the Ray framework to maximize **GPU utilization**, ensuring efficient training and inference of custom LLMs.
- Integrated models from **OpenAI, HuggingFace, Gemini, and Anthropic** to enhance the platform's AI capabilities.
- Successfully deployed the entire AI backend on **Amazon EC2** for robust performance.
- Developed custom metrics for scaling **Celery workers** via **CloudWatch**, improving resource efficiency.

- **GARUDA AI:**
- **Objective:** Developed an AI-driven blog writing engine designed to boost website rankings by generating SEO-optimized content.
- **Tech Stack:** Google Keyword Planner, Ahrefs, Generative AI, and LLMs.
- **Key Achievements:**
- Used **Google Keyword Planner** to analyze search volume and identify high-impact keywords for SEO content generation..
- Scrape data from targeted websites using **Ahrefs** to gather comprehensive and relevant information.
- Automatically generate five optimized blog titles, complete with keyword tags, to enhance **SEO performance**.
- **Produce high-quality**, SEO-friendly content based on the generated titles using advanced generative AI and **LLM models**.

F(x) Data Labs PVT LTD — Ahmedabad, Gujarat

MACHINE LEARNING ENGINEER | Jan 2022 – Dec 2023

- **The GMR Group - MCP Prediction (Time Series Forecasting):**
- **Objective:** Developed a predictive model for forecasting the **Market Clearing Price (MCP)** in the **Day-Ahead Market (DAM)** using a Random Forest algorithm, achieving an accuracy of **70% (±6%)**.
- **Key Achievements:**
- Expertly translated business requirements into precise **mathematical metrics** for robust analysis.
- Proficient in data **cleaning** and **transformation** to ensure high **data quality** and **reliability**.
- Built a strong foundation in **statistics and probability**, specializing in hypothesis testing, A/B testing, regression analysis, and Bayesian testing.
- Integrated data from diverse sources—including **IEX Market**, weather data, and Google Trends—using Selenium for web scraping, followed by comprehensive data cleaning and transformation.
- Designed a **confidence metrics** model to help traders assess the reliability of MCP forecasts for each time block.
- Employed **Monte Carlo simulation** to demonstrate that our model-based trading strategy achieved an **11% improvement** in profitability compared to random trading strategies.
- Proficient builds a wide range of machine learning techniques, such as **classification**, **clustering**, and **Genetic algorithm**.
- **Video Rendering Wowsly (Computer Vision):**
- **Objective:** Developed an API using **FastAPI** that allows multiple users to upload a **video** and a corresponding CSV file. Based on the CSV, the system automatically edits the video by overlaying text on specified frames using **OpenCV**, stores the results in an **S3 bucket**, and generates a tracking token for progress monitoring.
- **Key Achievements:**
- Edited videos by overlaying text on video clips at predetermined positions and durations based on CSV inputs.
- Utilized Docker & Docker Compose to manage microservices including FastAPI, Celery, Redis, and Flower.
- Designed an AWS-based asynchronous **video-processing architecture** using Lambda, EC2, S3, Celery, and Redis for scalable job processing and media storage.

EDUCATION & CERTIFICATIONS

L.D. College Of Engineering (GTU) — Ahmedabad, Gujarat

Bachelor of Engineering, Computer Engineering | 2018 – 2022

Cumulative CGPA: 8.62/10

Certifications & Training: Online Course in [Statistics with Python by Michigan University \(Coursera\)](#), [Mathematics for Machine Learning by Imperial College London](#), Computational Thinking and Data Science(MIT).

Languages: Native proficiency in Gujarati and Hindi; professional working proficiency in English.