

Ronak Kataria

Sr. Technical Lead - Data Science | Solutions Architect | AWS and Azure Certified

Solution-oriented and passionate Data Scientist and LLMops Architect with 7+ years of experience in Machine Learning, MLOps, and Generative AI. Proven expertise in designing scalable, secure, and production-grade AI/ML solutions across cloud ecosystems (AWS, Azure). Led high-impact initiatives, including RAG-based productivity tools, multi-agent LLM architectures, and enterprise LLMops platforms. Adept at driving cross-functional collaborations and delivering transformative AI outcomes.

✉ ronak.kataria1@gmail.com

☎ +91 - 8758682083

📍 Pune, India

🌐 [linkedin.com/in/katariaronak](https://www.linkedin.com/in/katariaronak)



TECHNICAL SKILLS

Amazon Web Services

Python

NodeJS

Machine Learning

Solutions Architecture

Generative AI

NLP

MLOps

TensorFlow

Large Language Models (LLMs)

Computer Vision

Agentic AI

OpenAI GPT

Neural Networks

GraphQL

Microsoft Azure

Google Cloud Platform

DataBricks

Snowflake

LLMOps

MLflow

MySQL

CI/CD

MongoDB



WORK EXPERIENCE

Senior Technical Lead – AI Engineering

KPIT Technologies (CTO AI Office)

10/2024 - Present

Pune, India

- Led the architecture and implementation of enterprise-grade scalable RAG-based productivity tools with LangChain and various LLMs (GPT, LLaMA, Qwen, Claude Sonnet, etc.), boosting employee productivity by 40%, reducing time to go-live, and increasing project revenues across KPIT.
- Headed the design and development of a LLMops accelerator platform, enabling smooth orchestration, fine-tuning, and deployment of LLMs in enterprise environments.
- Worked hands-on with LLM fine-tuning, multi-agent architectures, and diverse embedding models like Nomic, OpenAI, and Hugging Face.
- Delivered 30+ Agentic AI interventions in collaboration with cross-functional teams across automotive verticals—analyzing current workflows, identifying AI/GenAI automation opportunities, designing target-state architectures, and delivering working AI solutions.
- Led and mentored a team of 12, fostering collaboration, knowledge sharing, and delivery excellence.

Data Scientist

LTIMindtree

07/2022 - 10/2024

Pune, India

- Worked as an MLOps Solutions Architect and Data Scientist in the MLOps Center of Excellence (CoE), building a multi-cloud MLOps accelerator supporting AWS, Azure, GCP, and Databricks.
- Got company wide recognition for most viewed blog in 2023 Q4 on LLM Operationalization: <https://www.ltimindtree.com/blogs/embracing-the-future-with-generative-ai-operationalizing-large-language-models-using-snowflake-and-aws-sagemaker/>
- Architect and developed MVI for scalable Generative AI based LLMops platform for seamless implementation of RAGs and LangChain based pipelines for various Large Language Models (LLMs) like GPT, Llama2, T5, DistilBERT, Stable Diffusion, etc.
- Designed and developed RNN-based ML Algorithms for sales forecasting and leads generation for a leading bank client using KNN, LSTM, K-Means, and regression algorithms. Built MLOps pipelines and serverless inference functionality on AWS and data warehousing on Snowflake.
- Worked on an NLP Document AI using sentiment analysis, keyword spotting, topic modeling, K-Means clustering and tokenization algorithms.
- Received **Pat On The Back Award** for innovative thinking.

Software Engineer

Gateway Technolabs Pvt. Ltd. (Client - Limitless Technology, UK)

12/2020 - 07/2022

Ahmedabad, India

- Contributed heavily in architecting and developing a highly secure and scalable CRM application for UK-based Fortune 500 clients.
- Built fully serverless and largely decoupled production solutions by seamlessly integrating over 20+ AWS Services, including fully automated infrastructure deployment using AWS CloudFormation and Code Pipeline.
- Played a vital role as lead architect for Machine Learning based product features like Conversational AI, Task Dispatcher, NLP-based Sentiment Analysis of customer satisfaction, and self-learning custom task dispatcher algorithm.
- The AI-based decision-making features eliminated the need for manual configurations and human interference, which is a successful feature, that accelerated the overall product revenue growth of over 37% in 2021 Q4.
- Developed audio language conversion and closed-captioning tool for live stream video for Healthcare product, using NLP techniques and AWS.



WORK EXPERIENCE

Cloud and Machine Learning Engineer Volansys Technologies Pvt Ltd

07/2019 - 12/2020

Ahmedabad, India

- Developed end-to-end CI/CD pipelines and ML based product for object detection and tracking, as well as keywords spotting and NLP based POC.
- Developed OTA Portal and Data Science Dashboard, including Frontend, Backend, Database, Cloud, and DevOps integration on AWS.
- Rewards and Recognition: **Emerging Star Award** (Jan'20) and **Pat On The Back Award** [Twice] (Dec'19 and Sept'20)

Machine Learning Engineer SNDK Corp

02/2019 - 07/2019

Ahmedabad, India

- Worked on designing and optimizing custom Neural Network using TensorFlow, Keras and TFLite for low-powered devices.
- Design and develop the Alexa skill sets.
- Managed CI/CD pipelines and AWS infrastructure as a code using AWS CloudFront and Terraform templates.



EDUCATION

Bachelor of Technology - Electronics and Communications Engineering Charotar University of Science and Technology

08/2014 - 05/2018

CGPA 8.24



PORTFOLIO PROJECTS

Voice-Enabled Real-Time Interactive RAG Application

- Developed voice-enabled real-time interactive application using Retrieval-Augmented Generation (RAG) with LangChain, OpenAI Whisper, TTS-1 and GPT-4 APIs
- Implemented Python middleware and Flask-based RESTful APIs on AWS Lambda and API Gateway for processing speech input and managing backend functionalities.
- Utilized OpenAI Whisper API for converting user speech to text, and OpenAI TTS-1 API to convert text responses back to speech.
- Employed AWS CloudFront for content delivery, AWS Lambda for parallel processing, and AWS ElastiCache for caching to optimize interaction latency.
- Incorporated NLU techniques using AWS Comprehend for handling pauses, interruptions, and context management, and AWS Step Functions for orchestration.
- Deployed RAG pipeline using LangChain for document retrieval, augmentation, and response generation with OpenAI GPT-4. Utilized AWS SageMaker for model training and deployment.
- Technologies:** AWS Lambda, API Gateway, CloudFront, ElastiCache, S3, DynamoDB, SageMaker, Step Functions, WebRTC, Python, Flask, OpenAI Whisper, TTS-1, and GPT-4 APIs, LangChain, Pinecone.

IoT Device Management and Data Science Dashboard for HVAC and Water Heaters

- Designed and developed a One-Stop Data Science dashboard for managing and visualizing end-user HVAC and Water Heater devices.
- Designed and developed DevOps and Cloud workflow to deploy architecture, devices to cloud communication, manage devices and ETL services, visualize device data and generate big-data analytics.
- Technologies:** AWS (S3, EC2, RDS, ECS, ECR, Glue, Athena, CloudFormation, Load Balancer, Systems Manager, SNS, CloudFront, Route53, CloudWatch, Event Bridge), Docker, Apache Airflow, Apache PySpark, Python, Dash-Plotly, ReactJs, PostgreSQL.

Home-Office Automation and Surveillance System

- Designed and developed and train the Deep Learning and Computer Vision based Neural Network using TensorFlow and Python for running a neural network on low-power edge devices.
- Design and Implementation of REST API in NodeJs (Azure Functions), CI/CD pipelines, and cloud architecture for the device to communicate with Azure using Azure Core Services (IoT, API Management, Blob Storage, Functions, ML Studio, CosmosDB, Notification Hub, and Data Lake Analytics).
- Technologies:** Azure, Python, NodeJs, MQTT, Supervised Learning for ML, Computer Vision (OpenCV), TensorFlow, Keras, TFLite, Bash (Shell Scripting), NVIDIA Jetson Nano, Alexa developer tools.

Deep Learning based Generative Adversarial Network (GAN) to Generate Faces

- Developed and implemented Generator-Discriminator based Deep Convolutional GAN (DCGAN) architecture for generating realistic human faces.
- The network is trained on CelebA dataset containing more than 200k celebrity images.
- Implemented Generator-Discriminator based on Conv layer sequence of conv2D, leaky relu, sigmoid, batch normalization, dropout layers.
- Technologies:** PyTorch, Python, Deep Convolutional GAN (DCGAN), Convolutional Neural Network (CNN), Pandas, Numpy, Scikit Learn, Matplotlib.



CERTIFICATIONS

AWS Certified Solutions Architect - Associate

Amazon Web Services

AWS Certified Cloud Practitioner

Amazon Web Services

Microsoft Certified Azure Fundamentals

Microsoft

Deep Learning NanoDegree

Udacity (Facebook AI Labs and AWS)

AWS Certified Machine Learning Specialty

Amazon Web Services

Microsoft Certified Azure AI Fundamentals

Microsoft

Machine Learning NanoDegree

Udacity (Kaggle)

Generative Adversarial Networks (GANs) Specialization

deeplearning.ai (Coursera)



SOFT SKILLS

Client Communication

Critical Thinking

Team Player

Decision Making

Problem Solver

Rapid Learner



LANGUAGES

English

Full Professional Proficiency

Hindi

Full Professional Proficiency

Gujarati

Native or Bilingual Proficiency