

Ghanshyamsinh Zala

Ahmedabad, Gujarat, India | +91 9106688144 | ghanshyamsinhzala70@gmail.com
[LinkedIn](#) | [GitHub](#) | [Portfolio](#)

PROFESSIONAL SUMMARY

AI/ML Engineer focused on Generative AI, RAG systems, agentic workflows, and Python backend development. Currently working at Turing as an Agentic AI & Data Cloud Software Engineer, building production-grade LLM and retrieval services, data pipelines, cloud infrastructure, and deployment workflows. Previously worked at Mastek on enterprise RAG and LangGraph-based AI solutions, including a 67% reduction in query latency across 50+ internal workflows. Experienced in LLM applications, information retrieval, NLP, machine learning, APIs, data engineering, and production AI systems.

TECHNICAL SKILLS

Programming Languages: Python, SQL, JavaScript, Java, C#, Go.

Generative AI: RAG, LLMs, Prompt Engineering, Agentic AI, LangGraph, CrewAI, Llama 3.1, OpenAI APIs, Hugging Face, Ollama.

Machine Learning: PyTorch, TensorFlow, Scikit-learn, NLP, BERT, BART, Computer Vision, GANs, LSTM, SHAP.

Backend & Data: FastAPI, REST APIs, WebSockets, Pandas, PostgreSQL, MySQL, SQL Server, MongoDB.

Retrieval & AI Systems: FAISS, ChromaDB, Pinecone, BM25, Semantic Search, Hybrid Search, Embeddings.

Deployment & Tools: Docker, Kubernetes, MLflow, Azure AI Studio, Streamlit, Gradio, Git, GitHub.

Languages: English, Hindi, Gujarati.

PROFESSIONAL EXPERIENCE

Turing

Agentic AI & Data Cloud Software Engineer

Remote

Mar 2026 – Present

- Promoted from Agentic AI Trainee to full-time Software Engineer, developing production-grade RAG and LLM services integrated with APIs, databases, and enterprise data stores.
- Built end-to-end data engineering workflows, including ETL, streaming, validation, and data preparation for scalable machine learning and LLM systems.
- Implemented cloud infrastructure, infrastructure-as-code, and CI/CD workflows for secure and scalable AI model serving.
- Monitored model quality, latency, reliability, and system observability while collaborating with product and engineering teams.

Turing

Agentic AI Trainee

Remote

Feb 2026 – Mar 2026

- Built prototype RAG pipelines and integrated vector stores for task-specific LLM applications.
- Fine-tuned and evaluated LLM checkpoints using task-specific datasets and evaluation workflows.
- Contributed to ETL, data preparation, retrieval experiments, and model deployment workflows.

Mastek

Data Analytics & AI Engineer Intern

Ahmedabad, India

Aug 2025 – Jan 2026

- Built and optimized enterprise RAG workflows using Python, reducing query latency by 67% across 50+ internal workflows.
- Developed agentic workflows with Python and LangGraph to automate multi-step data analysis and decision-support processes.
- Worked on LLM-based solutions involving retrieval, prompt engineering, workflow orchestration, and enterprise data integration.
- Contributed to the integration and deployment of AI services within an enterprise environment while collaborating with cross-functional Agile teams.

EDUCATION

B.Tech in Computer Science & Engineering

Jan 2022 – Expected 2026

Darshan University

CGPA: 8.53/10

SELECTED PROJECTS

- AI Agentic Studio** | *Python, LangGraph, CrewAI, LLMs, RAG* 2026
- Built a multi-agent research engine orchestrating Planner, Researcher, Writer, Fact-Checker, and Quality-Gate agents to transform web data into technical whitepapers.
 - Implemented workflow orchestration, conditional routing, recursive quality loops, and agent-to-agent task handoffs for reliable multi-step execution.
 - Integrated LLM-based research, retrieval, verification, and local model inference to support privacy-aware document processing.
- Hybrid RAG Assistant** | *Python, FAISS, BM25, RAG, Streamlit* 2026
- Developed a multi-format RAG system supporting PDF, DOCX, TXT, HTML, and CSV documents with semantic and keyword retrieval.
 - Combined FAISS vector search with BM25 keyword retrieval to implement hybrid information retrieval for improved document grounding.
 - Integrated live web search and Streamlit-based interaction for document question answering and research workflows.
- Voice Assistant Pipeline** | *Python, JavaScript, FastAPI, WebSockets, Faster-Whisper, LLM* 2026
- Built a near-real-time conversational voice assistant connecting speech-to-text, LLM response generation, and text-to-speech through a full-stack pipeline.
 - Implemented WebSocket-based communication and Faster-Whisper speech recognition for low-latency conversational interaction.
 - Integrated asynchronous backend processing to coordinate audio streaming, inference, and response delivery.
- ByteBrief** | *Python, BERT, BART, Whisper, Hugging Face* 2026
- Developed a multimodal AI summarization platform combining BERT extractive summarization with BART abstractive summarization.
 - Integrated Whisper-based speech processing to extend summarization workflows to audio content.
 - Built modular NLP pipelines for context-aware summarization and sentiment-oriented text analysis.

ACHIEVEMENTS

- 1st Prize – National Code Carnival 2024:** National-level coding competition organized by Atmiya University and GSDC Group.
- 2nd Winner – Codearena Hackathon 2025:** Corporate-sponsored hackathon organized by Conversentech Company.
- 2nd Runner-Up – Half-Hackathon 2025:** Hackathon organized by RITA, B.H. Gardi College, and S.F.S. Edu. Expo.
- State-Level Winner – Sci-Pie Poster Competition:** Recognized at Darshan University for communicating scientific and technical concepts through visual design.
- Winner – G20 Summit Essay Writing Competition 2023:** Awarded by NSS Club, Darshan University, for critical thinking and communication on global topics.

LEADERSHIP & TEACHING

- Programming Club – Darshan University** Rajkot, India
Design Team Lead & Vice President Mar 2023 – May 2025
- Led a 20-member cross-functional team and helped organize technical events and hackathons for 500+ participants.
 - Mentored students on programming projects and contributed to technical community initiatives across the university.
- Darshan Institute of Engineering & Technology** Rajkot, India
Teaching Assistant Jan 2024 – Apr 2024
- Mentored 50 undergraduate students in front-end development, debugging, deployment, and project execution.

CERTIFICATIONS

- Oracle:** Fusion AI Agent Studio Certified Developer Professional – Rel 1 (2026).
- Microsoft:** Develop Your Own Custom Copilots with Azure AI Studio (2025).
- freeCodeCamp:** Machine Learning with Python (2024).
- HackerRank:** Certified in Go, C#, and Software Engineering.