

Nikhil Vankala

+91 8767029836 | nikhilvankala2004@gmail.com | [linkedin.com/in/nikhil-vankala](https://www.linkedin.com/in/nikhil-vankala) | github.com/Nikhil1081 | Bengaluru, India

PROFESSIONAL SUMMARY

Entry-level AI/ML engineer with internship experience in generative AI, LLMs, RAG, NLP, and deep learning. Proficient in Python, PyTorch, TensorFlow, and FastAPI. Skilled in building end-to-end AI pipelines, REST APIs, model deployment, and CI/CD workflows.

EDUCATION

Gandhi Institute of Technology and Management Bengaluru, India
B.Tech – Computer Science and Business Systems | CGPA: 7.3 2023 – 2027
– Relevant Coursework: Machine Learning, Deep Learning, Data Structures & Algorithms, OOP, Database Management

EXPERIENCE

Artificial Intelligence and Machine Learning Intern May 2026 – June 2026
JMR Infotech Bengaluru, India
– Engineered and deployed LLM-based AI pipelines for enterprise banking use cases, integrating Oracle OFSAA/FLEXCUBE via Python and REST APIs, cutting integration time by 40%.
– Built RAG systems using vector search and semantic retrieval, improving query accuracy by 35% across 500+ test cases.
– Executed UAT automation pipeline converting 300+ Excel test cases to JSON/CSV for LLM-based script processing, reducing manual effort by 50%.
– Collaborated cross-functionally to design end-to-end AI solutions using FastAPI, Docker, and CI/CD pipelines, accelerating delivery by 25%.

AI/ML Intern Jan 2026 – Feb 2026
IBM SkillsBuild via EduNet Foundation Remote
– Applied supervised and unsupervised ML models using scikit-learn and TensorFlow, achieving 88% average model accuracy across 5 datasets.
– Executed data preprocessing, feature engineering, and EDA on real-world datasets using Pandas and NumPy, reducing noise by 20%.
– Visualized model performance using Matplotlib and Seaborn; produced technical reports adopted by 3 project teams.

Generative AI Intern Jul 2025 – Oct 2025
Infosys Springboard Remote
– Designed and shipped 3 generative AI/NLP applications including a prompt engineering pipeline and an LLM text classifier, reducing response time by 30%.
– Optimized prompt workflows using advanced prompt engineering, improving output quality and cutting hallucination rate by 30%.
– Completed 200+ hours of AI/ML training covering transformer models, fine-tuning, deep learning, and responsible AI practices.

PROJECTS

Multimodal RAG Study Assistant | *Python, FastAPI, LangChain, FAISS, Vector Search* 2026
– Architected a multimodal RAG system over PDFs, diagrams, and text using FAISS vector DB and semantic embeddings; achieved sub-500ms query response time.
AI Voice Interview Platform | *Python, FastAPI, Speech-to-Text, NLP, Firebase, Railway* 2025
– Developed an async AI interview platform with speech-to-text pipelines and LLM-generated feedback; reduced processing latency by 40% after deployment on Railway.
AI Resume & Portfolio Builder | *Python, FastAPI, Docker, GitHub Actions, CI/CD* 2025
– Engineered an ATS-optimized resume generation tool using LLMs with automated PDF export; enforced CI/CD via GitHub Actions across 3 deployment platforms.

TECHNICAL SKILLS

Languages: Python, Java, C++, SQL, HTML, CSS
AI/ML Frameworks: TensorFlow, PyTorch, Keras, scikit-learn, HuggingFace Transformers, LangChain, NumPy, Pandas, Matplotlib, Seaborn
AI/ML Concepts: Machine Learning, Deep Learning, Neural Networks, NLP, Generative AI, LLMs, RAG, Prompt Engineering, Feature Engineering, Model Deployment, Fine-tuning, Transfer Learning
Tools & DevOps: FastAPI, Docker, Git, GitHub Actions, FAISS, Vector Databases, REST APIs, CI/CD, Agile/Scrum
Cloud: AWS, Google Cloud Run, Railway, Firebase, Streamlit

CERTIFICATIONS

IBM SkillsBuild: Advanced AI & Data Science (2026) · **Infosys Springboard:** Generative AI & LLMs (2025) · **EduNet Foundation:** Software Development & Automation (2026)