Vishal Singh Yadav

LinkedIn: linkedin.com/in/tokudai/ Email: vishal.singh@alumni.iith.ac.in
Github: github.com/tokudai-vs Mobile: +91-9555756235
Address: NCR / Hyderabad, India

Summary

Senior Data Scientist with 5+ years of experience spanning academic ML research and industry-scale production systems. Strong foundation in graph learning, representation learning, and probabilistic modeling, combined with hands-on delivery of ML and GenAI systems in healthcare, insurance, and manufacturing domains. Experienced in taking research-heavy ideas into production, including LLM-based systems, agentic workflows, forecasting, and real-time inference. Experience translating research-driven ML approaches into production systems under real-world engineering and operational constraints.

Experience

Tredence Inc. Hyderabad, India
Associate Manager (Senior Data Scientist) *Sep 2025 – Present*

- Working on applied ML and GenAI systems within the TMT domain, spanning forecasting, anomaly detection, and LLM-based analytics use cases.
- Contributing to analysis and modernization of ML and GenAI platforms through system profiling, code analysis, and performance evaluation.
- Supporting production ML pipelines and deployment workflows with a focus on reliability, latency, and production readiness.

Carelon Global Solutions (*Elevance Health*) Hyderabad, India
Associate Data Scientist III *Jul 2023 – Sep 2025*

- Built large-scale forecasting and risk models for the Cost of Care platform (60M+ members), delivering 3-month state-level utilization forecasts with Bayesian uncertainty quantification and reducing error variance by ~12%.
- Developed NLP- and probabilistic ML solutions across clinical, financial, and demographic data, improving underwriting accuracy, reducing turnaround time by 20%, increasing coverage from 48% to 81%, and enabling ~$1.8M in projected annual savings.
- Implemented LLM-powered systems including NL-to-SQL pipelines and a secure clinical Q&A assistant using LLAMA2, LangChain, and QLoRA-tuned RAG, achieving ~95% query accuracy and ~31% performance gains.
- Deployed and operated production ML systems across AWS, Azure, and on-prem environments using Dockerized services, CI/CD pipelines, and automated validation, monitoring, and retraining workflows.

Qulabs Software India Pvt Ltd Hyderabad, India
Machine Learning Engineer *Mar 2023 – Jul 2023*

- Built a semantic search and document QA system using BERT embeddings and vector search over 100K+ documents, enabling fast, context-aware retrieval without reliance on LLMs.
- Deployed NLP models for production use cases and mentored junior engineers on ML workflows, code quality, and deployment practices, improving maintainability and reuse across projects.

Krama Lab, IIT Hyderabad (*Machine Learning Research*) Hyderabad, India
Research Assistant *Jan 2020 – Dec 2022*

- Conducted applied research on graph neural networks, point cloud segmentation, and few-shot learning, developing model architectures for structured reasoning and sparse-data industrial settings.
- Designed and evaluated GNN-based approaches for knowledge graph completion and link prediction under low-resource constraints, focusing on representation learning and generalization.

Selected Consulting Experience

Independent Consulting (Part-time) Remote / India
Consultant, Machine Learning & GenAI Systems *Jul 2025 – Sep 2025*

- Contributed to the development of an agent-based ML system for industrial analytics, integrating natural-language interfaces with existing data, sensor, and optimization infrastructure.
- Built supporting components for agent orchestration, schema-aware query routing, and real-time data access.
- Performed system-level analysis to address latency, noisy data, and operational constraints, supporting production-readiness and reliability.

Education

- **Indian Institute of Technology Hyderabad** — Hyderabad, India
  M.Tech (by Research) *in Computer Science and Engineering GPA: 8.38 / 10* — *Jan 2020 – Dec 2022*
- **Guru Gobind Singh Indraprastha University** — New Delhi, India
  B.Tech *in Computer Science and Engineering Percentage: 71.2%* — *Jul 2014 – May 2018*

Skills

- **Generative AI, NLP & Machine Learning**: Large Language Models (LLMs), Prompt Engineering, Retrieval-Augmented Generation (RAG), AI Agents, Fine-tuning (LoRA, QLoRA), Transformers, BERT, spaCy, Semantic Search, Graph Neural Networks, Knowledge Graphs, Few-Shot Learning, Anomaly Detection, Time-Series Forecasting, PyTorch, TensorFlow, Scikit-learn, XGBoost
- **MLOps & Cloud**: AWS (EC2, S3, SageMaker, EKS), Docker, Kubernetes, CI/CD, MLflow, Weights & Biases, FastAPI, Model Serving, ONNX, Model Monitoring, Performance Tuning, A/B Testing
- **Data Engineering & Platforms**: PySpark, Snowflake, Distributed Computing, Feature Stores, ETL Pipelines, Time-Series Analysis, Vector Databases (FAISS)
- **Programming Languages**: Python, SQL, C++, Bash

Projects

- **CodeChakra: AI-Powered PR Review Assistant (OSS, WIP)**: Built an AI-assisted system to automate GitHub pull request reviews by combining static code analysis, security checks, and latency profiling with LLM-based reasoning for performance and reliability feedback. Designed a modular, agent-based architecture with encrypted I/O and CI/CD integration, enabling local, high-trust deployment using Docker.

Achievements

- **Go Above Award (x2) – Carelon Global Solutions**: Recognized twice for high-impact delivery under tight timelines, including rapid launch of a member insights dashboard and expedited execution of a critical analytics initiative.
- **Rapid Product Deployment – Carelon Global Solutions**: Recognized by senior leadership for fast-tracked development and production deployment of a key analytics product supporting business-critical decision making.
- **Mentor – AI and Advanced Technologies**: Provided technical mentorship to working professionals through the IIT Hyderabad & NSE Talentsprint program, covering ML workflows and production-grade AI system development.

Certifications

- **Advanced Certification in Generative AI – Carelon / Prizmato**: Production deployment of LLMs, QLoRA fine-tuning, RAG system design, and compliance-aware GenAI workflows.
- **NVIDIA Deep Learning Institute**: NLP and Deep Learning (Transformers), AI for Anomaly Detection and Data Engineering, Accelerated Computing and Data Science.

Publications

- **Context-Aware Question Routing in Community Question Answering Sites**: *Vishal Singh Yadav, Manish Singh* May 2023. Proposed a deep learning–based approach for intelligent question routing using contextual embeddings and semantic similarity to improve expert matching in community Q&A platforms.