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

Data Scientist (Associate DS 3) with 5+ years of combined experience — 3 years in applied research and 2+ in industry delivering scalable NLP, GenAI, and risk modeling systems. Built and deployed ML pipelines across LLMs (LLAMA, RAG, QLoRA), time-series forecasting, and NL2SQL — powering $1.5M+ in impact and 15x outreach growth. Deployed solutions across AWS, Azure, and hybrid infrastructure in healthcare and insurance use cases.

Experience

- **Carelon Global Solutions (*Elevance Health*)** Hyderabad, India
  Associate Data Scientist III *Jul 2023 – Present*
  - Designed and integrated forecasting modules with anomaly detection and GenAI augmentation into the Cost of Care platform, processing claims across 60M+ members into 3-month predictions — enabled state-level proactive interventions and accelerated action plans.
  - Applied NLP-driven feature extraction and optimisation over financial, clinical, demographic data — reduced underwriting time by 20% and improved real-time rating accuracy — projecting $1.5M cost savings by 2025.
  - Built probabilistic models to identify high-cost claimants and special conditions, driving $300K in annual savings through predictive cost analytics and improved risk visibility. Boosted condition identification coverage from 48% to 81%, enabling a 15x increase in targeted outreach.
  - Engineered a NL-to-SQL pipeline with fine-tuned LLMs, semantic vector search, and interactive query flow visualization — delivering 95% accuracy and outperforming prior solutions by 31%.
  - Developed a medical QA assistant using LLAMA2 and LangChain with RAG pipeline, fine-tuned via QLoRA (4-bit) for domain-specific precision — delivered a secure, scalable system for clinical question answering.
  - Automated manual ticket triage using a hybrid Word2Vec+BERT embedding model with XGBoost — eliminated repetitive classification work, saving 8–10 FTEs and standardising prioritisation workflows.
  - Achieved a 40% boost in underwriting decision throughput and $100K+ in cumulative savings by integrating ML-based risk scoring, rule-driven clinical flagging, and LLM-powered automation across core workflows.
  - Assisted model deployment across hybrid infrastructure (AWS, Azure, on-prem) using Docker-based pipelines — supported scalable rollout of real-time scoring APIs and dashboard integration.
- **Qulabs Software India Pvt Ltd** Hyderabad, India
  Machine Learning Engineer *Mar 2023 – Jul 2023*
  - Built a RAG-based chatbot from scratch for semantic document retrieval across 100K+ unstructured documents — enabled real-time question answering with domain specifc embedding and prompt tuning.
  - Mentored 3 interns and 3 junior developers in ML workflows, code quality practices, and model deployment — established standards that improved maintainability and review efficiency across the team.
- **Krama Lab, IIT Hyderabad (*Machine Learning Research*)** Hyderabad, India
  Research Assistant *Jan 2020 – Dec 2022*
  - **Industry Collaboration (GreatFour Systems):** Conducted applied research on point cloud segmentation, graph neural networks (GNNs), and few-shot learning techniques for real-world industrial use — contributed core model architectures for structured reasoning and sparse data classification in enterprise settings.
  - **Academic Research:** Designed and evaluated experimental GNN-based architectures for knowledge graph completion, low-resource link prediction, and structured representation learning.

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**: LLAMA, QLoRA, Fine-tuning, Prompt Engineering, RAG (Retrieval Augmented Generation), AI Agents, MCP (Model Context Protocol), Transformers, BERT, spaCy, Semantic Search, PyTorch, TensorFlow, Scikit-learn, XGBoost, Graph Neural Network, Knowledge Graphs, Few-shot Learning, Anomaly Detection
- **MLOps & Cloud**: AWS (EC2, S3, SageMaker), Docker, CI/CD, MLflow, Weights & Biases, FastAPI, Model Serving, ONNX, Performance Tuning, A/B Testing, Streamlit, Gradio
- **Data Engineering & Tools**: PySpark, Distributed Computing, Snowflake, Feature Stores, ETL Pipelines, FAISS, Time-Series Analysis, Data Augmentation
- **Programming Languages**: Python, SQL, C++ (STL, Boost), Bash

Projects

- **CodeChakra: Autonomous AI PR Review Agent (WIP)**: Building a secure, end-to-end AI agent that automates pull request reviews by simulating a full-stack engineering audit. Triggered on each GitHub PR, it performs static analysis, security scans, latency profiling (I/O, DB, network ops), and LLM-based suggestions for performance and resilience improvements. Architected for high-trust environments with planned homomorphic encryption and encrypted I/O. Open-source and extensible by design — PoC in progress to enable future community-driven modules and integrations.

Achievements

- **Rapid Product Deployment – Carelon Global Solutions**: Recognized by the Director of Data Science for leading the accelerated delivery of a critical product, enabling faster business insights and smoother deployment into production.
- **Go Above Award (x2) – Carelon Global Solutions**: Twice awarded for high-impact execution under pressure — launched a member insights dashboard within 24 hours and led the expedited delivery of the ARROW project under stringent timelines.
- **Mentor – AI and Advanced Technologies**: Delivered technical mentorship under the IIT Hyderabad NSE Talentsprint program — trained professionals in core ML workflows and production-grade GenAI systems.

Certifications

- **Advanced Certification in Generative AI – Carelon / Prizmato**: Covered production deployment of LLMs, QLoRA fine-tuning, RAG system design, HIPAA-compliant prompt engineering, and GenAI optimisation techniques.
- **NVIDIA Deep Learning Institute**: **NLP & Deep Learning:** Fundamentals of Deep Learning, Transformer-Based NLP Applications
  AI for Ops: AI for Anomaly Detection, Accelerated Data Engineering Pipelines
  Accelerated Computing: Fundamentals of 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 framework for intelligent question routing using contextual embeddings and semantic similarity scoring for improved query-to-expert matching.