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 with 5+ years of experience — 3 in applied research and 2+ delivering production-grade ML solutions in NLP, GenAI, and risk modelling. Skilled across the ML lifecycle, from prototyping to deployment, with projects spanning LLMs (LLAMA, RAG, QLoRA), forecasting, and NL-to-SQL systems. Focused on healthcare and insurance, with hands-on delivery across AWS, Azure, and hybrid cloud environments.

Experience

- **Carelon Global Solutions (*Elevance Health*)** Hyderabad, India
  Associate Data Scientist III *Jul 2023 – Present*
  - Developed & deployed forecasting modules with anomaly detection and GenAI for the Cost of Care platform, converting claims data (60M+ members) into 3-month state-level predictions to support proactive interventions.
  - Designed Bayesian hierarchical models to quantify uncertainty in cost-of-care forecasts, producing calibrated credible-interval outputs that improved pricing decisions and cut forecast error variance by ~12%.
  - Worked on NLP-based feature engineering and optimisation using financial, clinical, and demographic data to help reduce underwriting turnaround by 20% and improved real-time rating accuracy, contributing to projected $1.5M in projected cost savings by 2025.
  - Built probabilistic models to flag high-cost claimants and special medical conditions, enabling ~$300K in annual savings through improved cost visibility and risk stratification. Increased condition identification coverage from 48% to 81%, supporting a 15x expansion in outreach.
  - Created an NL-to-SQL pipeline using LLMs, semantic search, and visual query flows. Achieved ~95% accuracy gain and improved performance by ~31% over existing internal methods.
  - Co-developed a clinical QA assistant using LLAMA2, LangChain, and a QLoRA-tuned RAG pipeline, enabling secure, domain-specific medical question answering at scale.
  - Automated manual ticket triage workflows using statistical classifiers (Naive Bayes, XGBoost) with Word2Vec and BERT embeddings, used probabilistic baselines to guide model iteration and achieved effort savings of 8–10 FTEs across support teams.
  - Contributed to a ~40% improvement in underwriting throughput by building a clinical rules engine, integrating ML-based risk scoring, & enabling LLM-assisted workflow automation, delivering $100K+ in operational savings.
  - Supported deployment of ML models across AWS, Azure, and on-prem environments using Docker pipelines, ensuring stable rollout of real-time APIs and dashboards.
- **Qulabs Software India Pvt Ltd** Hyderabad, India
  Machine Learning Engineer *Mar 2023 – Jul 2023*
  - Built a RAG-style semantic search QA chatbot using BERT embeddings and vector search over 100K+ files, enabling fast, context-relevant answers via semantic similarity without LLMs.
  - Mentored 3 interns and 3 junior developers on ML workflows, code quality, and model deployment, improved maintainability and established reusable standards across the team.
- **Krama Lab, IIT Hyderabad (*Machine Learning Research*)** Hyderabad, India
  Research Assistant *Jan 2020 – Dec 2022*
  - **Industry Collaboration (GreatFour Systems):** Conducted independent research on point cloud segmentation, graph neural networks (GNNs), and few-shot learning under faculty guidance, developed core model architectures for structured reasoning and sparse data classification in industrial settings.
  - **Academic Research:** Designed and evaluated experimental GNN-based models for knowledge graph completion, link prediction, and representation learning under low-resource constraints.

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

Projects

- **CodeChakra: AI-Powered PR Review Assistant (OSS, WIP)**: Building an end-to-end AI assistant to automate GitHub PR reviews by simulating full-stack audits. It runs static analysis, security scans, and latency profiling for I/O, DB, and network ops, then uses LLMs for performance and reliability suggestions. Designed for high-trust environments with encrypted I/O, modular agent orchestration, and CI/CD integration, the Docker-based PoC supports local deployment. The roadmap includes features like ML-powered explainability, anomaly detection, and repo-wide insights.

Achievements

- **Go Above Award (x2) – Carelon Global Solutions**: Twice awarded for high-impact execution under critical timelines; launched a member insights dashboard within 24 hours and led the expedited delivery of the ARROW project during peak business pressure.
- **Rapid Product Deployment – Carelon Global Solutions**: Personally awarded by the Director of Data Science for leading the fast-tracked build and deployment of a key analytics product, accelerated insight delivery and enabling smooth production rollout.
- **Mentor – AI and Advanced Technologies**: Delivered technical mentorship to working professionals through the IIT Hyderabad & NSE Talentsprint program, trained cohorts in ML workflows and production-grade AI system development.

Certifications

- **Advanced Certification in Generative AI – Carelon / Prizmato**: Covered production deployment of LLMs, QLoRA fine-tuning, RAG system design, HIPAA-compliant prompt engineering, and optimisation techniques for GenAI workflows.
- **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 to improve query-to-expert mapping.