# Vishal Singh Yadav

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

## Summary

Data Scientist and ML Engineer with 5 years of combined experience — 2+ years in industry delivering production grade NLP and GenAI systems, and 3 years of research in GNNs, knowledge graphs and recommendation systems. Skilled in LLMs (LLAMA2, RAG, QLoRA), predictive modelling, and deploying ML solutions on AWS/Azure with proven business impact — including $1.5M in projected savings and 15x outreach improvement.

## Experience

- **Carelon Global Solutions (*formerly Legato Health Technologies*)** Hyderabad, India
  *Associate Data Scientist* *Jul 2023 – Present*
  - Designed and integrated AI-driven forecasting modules into the Cost of Care platform using anomaly detection and GenAI, transforming retrospective claims data into proactive, state-wise 3-month predictions — accelerating decision-making and enabling timely healthcare interventions.
  - Applied NLP, ML, and large-scale optimization to automate risk scoring using financial, clinical, and demographic data — reducing underwriting time by 20% and improving rating accuracy with live scoring and deployment pipelines, projected to save $1.5M 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 production-ready NL-to-SQL pipeline with fine-tuned LLMs, semantic vector search, and interactive knowledge graph generation — delivering 95% accuracy and outperforming prior solutions by 31%.
  - Built a medical QA system using LLAMA2 + RAG (LangChain), fine-tuned with QLoRA (4-bit) for domain-specific QA — improving context relevance and response quality.
  - Developed a support ticket classifier using Word2Vec+BERT embeddings with XGBoost, leveraging data augmentation and sampling to automate triage workflows and save 8–10 FTEs.
  - Automated underwriting workflows using ML and LLM-powered tools, integrating explainable risk scoring and clinical flagging — enhancing underwriter decisions and delivering $100K in impact with a 40% gain in efficiency.
  - Deployed ML models across hybrid infrastructure (AWS, Azure, on-prem) using Docker-based cloud pipelines, integrating business rules and dashboards to enable scalable, low latency inference and optimized model performance.
- **Qulabs Software India Pvt Ltd** Hyderabad, India
  *Machine Learning Engineer* *Mar 2023 – Jul 2023*
  - Built a domain-specific RAG-based chatbot for real-time document search and semantic matching over large unstructured datasets, enabling faster query resolution.
  - Guided junior developers on ML workflows, including data cleaning, model training, and deployment, improving overall code quality and delivery speed.
- **Krama Lab, IIT Hyderabad (*Machine Learning Research*)** Hyderabad, India
  *Research Assistant* *Jan 2020 – Dec 2022*
  - Conducted applied research on point cloud segmentation, graph neural networks, and knowledge graph completion under an industry-sponsored collaboration (GreatFour Systems).
  - Developed few-shot learning techniques for point cloud classification and designed GNN-based architectures for structured knowledge representation and reasoning tasks.

## Education

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

## Skills

- **Core ML**: PyTorch, TensorFlow, Scikit-learn, XGBoost, Keras, spaCy, Transformers, Hyperparameter Tuning, ONNX / TorchScript
- **Generative AI**: LLAMA2, QLoRA, RAG Architectures, LangChain, Prompt Engineering, Fine-Tuning, ChromaDB, Faiss, Pinecone, LlamaIndex
- **MLOps & Cloud**: AWS (EC2, S3), Docker, Kubernetes, CI/CD Pipelines, Git, FastAPI, Streamlit, Gradio
- **Data Engineering**: PySpark, Snowflake, ETL Pipelines, Distributed Computing, Feature Stores, Time-Series Databases
- **Advanced ML**: Graph Neural Networks, Few-Shot Learning, Anomaly Detection, Knowledge Graphs, AI Agents
- **Production ML**: Model Serving, A/B Testing, Performance Optimization, MLflow, Weights & Biases
- **Programming**: Python (NumPy, Pandas), C++ (STL, Boost), SQL, Bash
- **Tools & Platforms**: Linux, Jupyter, LaTeX, Matplotlib, Seaborn

## Achievements

- **Rapid Product Deployment, Carelon Global Solutions**: Recognised by the Director of Data Science for leading the accelerated deployment of a critical product, improving operational efficiency and enabling faster business insights
- **Go Above Award – Carelon Global Solutions**: Twice honoured for exceptional execution under pressure: delivered a member dashboard within 24 hours and expedited the ARROW project under stringent deadlines.
- **Mentor – AI and Advanced Technologies**: Delivered technical mentorship under an IIT Hyderabad and NSE Talentsprint initiative, guiding professionals in machine learning and GenAI fundamentals.

## Certifications

- **Advanced Certification in Generative AI (Carelon/Prizmato)**: Production deployment of LLMs, QLoRA fine-tuning methodologies, RAG system design, HIPAA-compliant prompt engineering, GenAI optimization 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.* Developed a method for improving question routing on community-driven Q&A platforms using context-based embeddings.