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 and knowledge graphs. Skilled in LLMs (LLAMA2, RAG, QLoRA), predictive modeling, and deploying ML solutions on AWS/Azure with proven business impact.

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 — enabling faster decision-making and timely healthcare interventions.
  - Engineered high-accuracy time-series forecasting using SARIMA and Prophet, reducing RMSE to 9.05 and unlocking early trend detection for smarter resource planning and cost control.
  - 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 real-time model deployment, projected to save $1.5M by 2025.
  - Built a medical QA system using LLAMA2 + RAG (LangChain), fine-tuned with QLoRA (4-bit) for domain-specific QA, significantly improving context relevance and response quality.
  - 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%.
  - Engineered a ticket classification pipeline using Word2Vec+BERT embeddings with XGBoost, leveraging data augmentation and sampling to automate workflows and save 8–10 FTEs.
  - Created interpretable risk-scoring and clinical flagging models with explainable AI techniques, aiding underwriters and delivering $100K in business impact.
  - Automated underwriting workflows using ML and LLM-powered conversational tools, reducing manual effort and enhancing operational efficiency by 40%
  - Deployed models across hybrid infrastructure (AWS, Azure, on-prem), integrating with business rules and dashboards to support real-time decisions.
  - Applied large-scale preprocessing and optimisation techniques to boost model performance, using Docker-based cloud deployments for scalability.
  - Built probabilistic models to identify recurring high-cost claimants and special conditions, enabling predictive cost trajectory analytics and driving $300K in annual savings through improved risk visibility and decision-making.
- **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, XGBoost, Transformers, spaCy, Keras, Scikit-learn, Hyperparameter Optimization
- **Large Language Models**: LLAMA2, RAG Architectures, QLoRA, LangChain, Prompt Engineering, Fine-Tuning
- **Cloud & MLOps**: AWS (EC2/S3), Docker, Kubernetes, Git, CI/CD Pipelines
- **Data Engineering**: PySpark, Snowflake, Distributed Computing, ETL, Feature Stores, Time-Series Databases
- **Specialized ML**: Graph Neural Networks, Few-Shot Learning, Anomaly Detection, Knowledge Graphs, AI Agents
- **Languages**: Python (NumPy/Pandas), C++ (STL/Boost), SQL, Bash
- **Tools**: Linux, Jupyter, LaTeX, Matplotlib, Seaborn, Snowflake

Achievements

- **Award for 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.