KOMAL NIRAULA

New York, NY | (347) 988-3775 | kn2505@nyu.edu | linkedin.com/in/komal-niraula/ | komalniraula.github.io/

SUMMARY

Machine learning engineer with experience in applied AI and cloud native ML systems, currently pursuing graduate studies in Computer Engineering. Skilled in building, evaluating, and deploying predictive models and LLM services using scalable pipelines, distributed systems, and cloud infrastructure. Experienced in time-series and predictive modeling, online experimentation, and microservice based ML systems that translate data into reliable forecasts. Particularly interested in efficient, production grade AI, with a focus on resource aware learning, scalable inference, multi-agent orchestration, and cloud native development of ML solutions.

EDUCATION

New York University, New York City, NY — Expected May 2026
Master of Science in Computer Engineering, GPA: 3.704

- **Coursework**: Machine Learning, Efficient AI Computing, Reinforcement Learning & Optimal Control, Deep Learning, Computing System Architecture, Applied Matrix Theory, Robo Advisors & Systematic Trading
- **Guided Study**: Network Analysis & Stochastic Modeling (under Prof. Shivendra Panwar)
- **Advanced Project**: Multi-Agent Debate System for Decision Making (under Prof. Yong Liu)
- **Achievements:** Winner- NYU Tech Duels 2025; presented arguments against strict tech regulation
- **Certifications:** NYU Tandon Data Science Bootcamp, IBM Dev Day: AI Demystified, Bloomberg Finance Fundamentals

University of Wolverhampton, Herald College, Kathmandu, Nepal — March 2022 – Feb 2024
Master of Business Administration, Grade: Distinction

- Full Scholarship from ING Group, Nepal

TECHNICAL SKILLS

Programming & Tools: Python, Java, SQL, Pandas, NumPy, Scikit-Learn, PyTorch, Git, FastAPI, Microservices, Distributed systems, Spark, Git, AWS (EC2, S3, Lambda, ECR, CloudWatch), Vector Search (FAISS, Pinecone)
Machine Learning: Linear & Logistic Regression, Clustering, Feature Engineering, Predictive Modeling, Sentiment analysis, Natural Language Processing, Transformer Models, CNN, SHAP, Optimization algorithms, Data Analysis & Visualization, Gradient boosting
Efficient AI & Model Optimization: Pruning, Quantization, Distillation, Low Rank Decomposition, NAS, LLM Accelerators
Data Pipelines & Large-Scale Systems: Distributed data processing (Spark), Feature pipelines, Batch & streaming workflows, Data validation, Pipeline orchestration, Scalable ML systems, Cloud data infrastructure
Model Evaluation & Experimentation: Forecast accuracy metrics (RMSE, MAPE), Precision, Recall, F1-score, Confusion Matrix analysis, Backtesting, Out-of-sample validation, Online experimentation, A/B testing, Model monitoring, Drift detection

PROJECTS

Adaptive Inference: Dynamic Efficiency Strategies for Large Language Model Inference — December 2025

- Engineered and evaluated early exit strategies for GPT-2, achieving 74-81% reduction in inference latency for classification task.

Runner-up, Alphathon 2025, Society of Quantitative Analyst — October 2025

- Applied NLP and quantitative modeling to detect and trade on narrative shifts in corporate storytelling; selected as a Top 3 out of 62 teams for innovative LLM driven financial signal research, presenting at the 'SQA Agentic AI in Investment' Conference.

Finetuning RoBERTa with AdaLoRA on AG News — March 2025

- Implemented parameter-efficient fine-tuning of RoBERTa using AdaLoRA on the AG News dataset, reducing trainable parameters by ~90% while maintaining competitive text classification performance.

EXPERIENCE

Graduate Research Assistant, **NYU Stern School of Business,** New York, NY — March 2025 – September 2025

- Researched predictive variables for growth and value portfolios of Fama French three factor model using 60+ years of data, enabling systematic analysis of how firm fundamentals and earnings announcement behavior drive factor returns across market regimes.
- Performed large-scale data cleaning, merging, and analysis on CRSP, Compustat, and IBES datasets to replicate and assess prior academic research, identifying where published growth-value results hold, weaken, or break across market periods.
- Modeled an OLS regression model to predict the earnings announcement timings, achieving 60% accuracy in identifying correct announcement weeks using rolling volatility of prior earnings surprises and annualized returns, with a fiscal year adjustment.

AI Project Manager, **Jyoti Bikash Bank,** Kathmandu, Nepal — December 2021 – September 2024

- Designed and deployed an in-house RAG system using Amazon Bedrock and LLaMA3, to answer in-house policy-related queries effectively, cutting branch policy inquiries to head office by 60%.
- Reconciled over $1 million in unsettled transactions by developing a reconciliation system that detects discrepancies between the payment gateway and the bank's system, flagging them for manual review.
- Built an automated presentation generation system using NLP techniques (KeyBERT, sentence embeddings, cosine similarity) to convert documents into PowerPoint slides, reducing manual slide creation time by 50%.
- Implemented Tableau dashboards to surface customer data quality issues (missing values, incorrect entries, inconsistencies), driving branch level remediation within a 1 week SLA and improving overall data reliability for reporting.