Koushik Vennalakanti

Bethlehem, PA | (484) 935-5320 | kov225@lehigh.edu
GitHub: github.com/kov225 | **LinkedIn:** linkedin.com/in/koushik-vennalakanti

Education

- **Lehigh University, Bethlehem, PA** **Jun 2025 – Dec 2026**
  M.S. in Data Science; GPA: 3.94 / 4.0
- **Birla Institute of Technology & Science (BITS) Pilani, India** **Aug 2020 – Jul 2024**
  B.E. in Mechanical Engineering; Minor in Data Science; GPA: 3.5 / 4.0

Technical Skills

- **Languages:** Python, SQL, R, C++
- **AI/ML:** PyTorch, scikit-learn, XGBoost, LightGBM, HuggingFace, LangChain, FAISS, ChromaDB, OpenAI API, SHAP, Pandas, NumPy, SciPy
- **LLMs & GenAI:** GPT-4o-mini, Mistral-7B, Diffusion Transformers, RAG, Fine-Tuning, LoRA, Prompt Engineering, Agentic Workflows, LLM Evaluation
- **Infrastructure:** FastAPI, Docker, Kafka, Redis, PostgreSQL, MLflow, Prometheus, Grafana, Streamlit, AWS (S3, EC2), Git, GitHub, CI/CD

Professional Experience

- **Research Assistant – Machine Learning Engineering** **Sep 2025 – Present**
  Lehigh University, Bethlehem, PA
  - Architected a config-driven ML pipeline in Python (scikit-learn, pandas, NumPy) with 16 modular components across 2,100+ lines, using abstract base classes and runtime class resolution via importlib so new extractors, models, and metrics plug in through JSON config without modifying orchestration code.
  - Implemented a time-series-aware training harness that retrains classifiers on rolling windows across 30+ years of daily data, automating hyperparameter search over 1,500+ combinations per fold against a custom objective, eliminating data leakage through causal feature construction and strict temporal splits.
  - Developed evaluation framework computing 30+ metrics with 100,000-iteration Monte Carlo simulation, four baseline models, and statistical significance testing (KS, chi-square, Wald-Wolfowitz), plus an automated weighted-vote comparator for model selection with full reproducibility.
- **Data Science Intern** **Jul 2023 – Feb 2024**
  CASHe (Fintech / Digital Lending), India
  - Built a hybrid company-search system using fuzzy string matching (TF-IDF n-gram) for known entities and GPT-based semantic resolution for unseen names, improving search accuracy by **35%**.
  - Enriched 500K+ customer location records via Google Maps API; applied geo-clustering to surface spatial demand patterns across 12 metro regions.
  - Engineered cluster membership as a predictive feature in the loan propensity model, strengthening regional risk differentiation and improving downstream model AUC.

Projects

- **Credit Intelligence Platform: LLM-Explained ML Serving Stack** **2026**
  Python, FastAPI, Docker, MLflow, Kafka, Redis, PostgreSQL, GPT-4o-mini, Mistral-7B, SHAP, XGBoost, LightGBM, Prometheus, Grafana
  - Built a 9-service ML inference platform on Docker Compose serving credit risk predictions and LLM-generated ECOA/FCRA-compliant adverse action explanations via SHAP TreeExplainer with configurable GPT-4o-mini and self-hosted Mistral-7B backends.
  - Trained a stacking ensemble (XGBoost + LightGBM + LR, 5-fold OOF meta-learner) on 2.2M records joined to 5 FRED macro series under point-in-time join, reaching **0.79 AUC** with Evidently PSI/KS drift monitoring and AUC-gated MLflow promotion.
- **Controllable Music Generation via Latent Space Navigation in Diffusion Transformers** **2025**
  Python, PyTorch, Streamlit, FastAPI, HuggingFace
  - Engineered an end-to-end generative audio pipeline integrating a 2B-parameter Diffusion Transformer with a 0.6B Qwen LM planner to synthesize 48kHz stereo audio, achieving a top-5 CLAP similarity of 0.528 vs. baseline ~0.125.
  - Built distributed inference via FastAPI microservices across heterogeneous hardware (CUDA + MPS), reducing generation time from ~3 min to ~1 min with a multi-page Streamlit dashboard for fine-grained controls.
- **Agentic RAG Pipeline with Tool-Use and Evaluation Harness** **2026**
  Python, LangChain, LangGraph, ChromaDB, OpenAI API, FastAPI, pytest
  - Built an agentic RAG system using LangGraph for multi-step reasoning with tool-use orchestration, ChromaDB vector store with OpenAI embeddings, and a routing layer that decomposes complex queries into sub-tasks before synthesizing answers.
  - Developed evaluation harness measuring retrieval precision, faithfulness, hallucination rate, and latency across 200+ test queries; deployed as FastAPI service with 15 pytest integration tests.