Chetan Chhetri

Connecticut cchhetri@my.bridgeport.edu ☎ 475-393-8383

in chetanchhetri Chetan-chhetri

About me

AI Engineer with more than 3 years of experience with detail-oriented Machine Learning Engineer focused on generative AI and language models, bringing deep knowledge of Python, prompt engineering, model fine-tuning, and serverless architectures. Experienced in implementing complex data pipelines, bias and fairness evaluation, and utilizing version control tools like Git. Recognized for cross-functional teamwork, clear documentation, and a continuous learning mindset aligned with emerging AI technologies and best practices.

Projects

- **Motion Detector – Agentic AI System** : - I spearheaded a research initiative focused on cognitive human-object interaction reasoning, using the Bongard-HOI and HAKE datasets. Here, I engineered and fine-tuned deep learning models to bridge the gap between synthetic and real-world understanding, raising model accuracy from 55% to 80% on more than 2,000 annotated image pairs through the integration of DVRL, attention mechanisms, and componential analysis. I implemented self-supervised representation learning, advanced data augmentation, and interpretability workflows to uncover and address reasoning failures, contributing key insights to cognitive approaches in computer vision
- **Personal Attorney – Legal RAG Assistant** : I architected a Legal Retrieval-Augmented Generation (RAG) assistant, employing ChromaDB, LangChain, and DeepSeek 8B (LoRA fine-tuning) to provide state-specific DMV and legal support. By deploying scalable REST APIs using FastAPI on AWS, I reduced retraining costs by 70%, maintained 98% information retrieval accuracy, and supported over 1,500 user queries per month across 50+ workflows. The project featured end-to-end prompt engineering, vector database optimization, OpenAPI-driven documentation, and robust MLOps for rapid iterative releases.
- **SmartMed AI – Pill Detection App** : I developed SmartMed AI, a robust pill detection application for pharmacies. Utilizing YOLO (v5–v12), ResNet, advanced multi-scale object detection, attention modules, and tracking, the system raised detection accuracy from 76% to 91%, processing more than 10,000 images monthly and improving operational reliability by 25%. The project included containerized deployment using Docker and FastAPI on AWS, as well as implementation of explainability tools like GradCAM and albumentations.
- **Sanskrit Transformer** : During Mar 2024 to Sep 2025, I led the development of an optimized Sanskrit Transformer for deep learning inference. I built and trained a Transformer-based NLP model using a custom 25,000-token vocabulary on ancient Sanskrit Vedic texts, incorporating advanced tokenization and architectural modifications. Leveraging NVIDIA GPUs, CUDA, cuDNN, and TensorRT, I reduced inference latency by 46% and boosted throughput by 39%, while profiling and tuning critical matrix operations for high efficiency. Through kernel-level performance optimization and benchmarking, I delivered a production-grade, real-time serving pipeline with sub-75ms inference times, ensuring the model's suitability for large-scale language applications.
- **Time Series Forecasting** : Since Sep 2025, I have been designing and evaluating time-series forecasting solutions tailored to highly volatile financial data. My contributions include building and benchmarking ARIMA, Meta Prophet, and XGBoost ensemble models, enhancing accuracy from 68% to 84% while reducing error by 31% during market swings. The solution features automated feature engineering, scalable model deployment tracked with MLflow, and in-depth reliability analysis using prediction intervals across 100,000+ monthly data points.

Education

MSCs **University of Bridgeport** 2024 – Present

- Coursework: Machine Learning & Deep Learning: TensorFlow, Pandas, YOLO, LangChain Web Development: Flask, FastAPI Data Handling & Processing: Pandas, SQLAlchemy LLM Applications: LangChain, OpenAI API

BTech **JNTUH** 2017-2021

- Digital Logic Design, VLSI Systems, Microprocessors, VHDL Programming, Em- bedded Systems, Computer Architecture

Experience

University of Bridgeport, E-Commerce and Digital Operations Associate — Bridgeport, CT

- Managed bookstore inventory and order placements using digital cataloging systems
- Utilized SQL-based tools and spreadsheets to track stock levels and process customer requests efficiently.
- Streamlined order workflows by creating structured logs, improving accuracy and reducing errors.
- **Tech stack**: College inventory management software, USPS tracking systems, SQL databases, and spreadsheet tools.

VLSI First, RTL Design Engineer — Hyderabad, India
2022–2023

- Designed and implemented digital systems using combinational and sequential circuits in VHDL, focusing on performance, reliability, and scalability.
- Verified designs using UVM test benches, constraints, and functional coverage analysis to ensure correctness and robustness.
- Collaborated with cross-functional teams to integrate modules and optimize the verification process.
- **Tech stack**: Questasim, Xilinx, VHDL, UVM.

Technologies

Languages: Python, Java, C/C++, JavaScript, SQL, NoSQL

Frameworks & Libraries: PyTorch, TensorFlow, Keras, scikit-learn, Hugging Face Transformers, LangChain, FastAPI, Flask, Pandas, NumPy, SQLAlchemy, YOLO, OpenCV, DVC

Databases & Cloud Deployment: MySQL, PostgreSQL, ChromaDB, AWS (S3, EC2, Lambda, SageMaker, Bedrock, Fargate), Azure ML, Azure OpenAI, GCP AI/ML, Google Cloud, Vertex AI, MLflow, Docker, Kubernetes, CI/CD automation, Serverless, REST APIs, Microservices, Cloud deployment, MLOps

AI / LLMs & RAG Systems: Large Language Models (GPT, LLaMA, Falcon, DeepSeek), Retrieval-Augmented Generation (RAG) systems, Generative AI, Prompt Engineering, Synthetic Data Generation, Named Entity Recognition, Sentiment Analysis, Text Generation, Model Fine-Tuning (LoRA), Model Optimization, Deep Learning Architectures (Transformers, CNNs), Model Evaluation & Interpretability, Feature Engineering, ARIMA, Prophet, XGBoost

Tools & Platforms: Git, GitHub, MLflow, TensorRT, CUDA, cuDNN, Postman, PyCharm IDE, ITSM tools, Performance Profiling, Benchmarking, System Optimization, Google Colab

Professional Strengths: Problem Solving, Debugging, Clean Code, Object-Oriented Programming, Design Patterns, Code Reviews, Agile & Scrum, Project Ownership, Cross-functional Collaboration, Backlog Management, Production Monitoring, Research, Attention to Detail

Strenghts

Problem Solving, Analytical Thinking, Cross-functional Collaboration, Communication Skills, Teamwork, Project Management, Continuous Learning

Certifications

Udemy: Game Developing with Spring Boot for Game API : Provided APIs for game clients to interact with, such as player profiles, leaderboards, and achievements.

The Complete Networking Fundamentals with CCNA: Covered OSI model, IP addressing, routing, switching, subnetting, and CCNA certification preparation.

VSD - Physical Design Flow: Covered timing closure, floorplanning, placement, routing, and hands-on transition from RTL to GDSII.