Stuti Bimali

West Haven, CT | (203) 298-2102 | bimalistuti@gmail.com | LinkedIn: linkedin.com/in/stuti-bimali | GitHub: github.com/stutibimali

PROFESSIONAL SUMMARY

AI/ML Engineer and M.S. Data Science graduate with 2 years of hands-on experience across internships and full-time roles building production-grade RAG pipelines and scalable AI systems. Experienced in fine-tuning and evaluating LLMs with domain-specific tasks.

TECHNICAL SKILLS

Programming Languages: Python, SQL, Java, R
ML/AI: NLP, Computer Vision, LLMs, RAG, Transfer Learning, Fine-tuning, PyTorch, Scikit-learn, Pandas, NumPy, TensorFlow
Cloud/Data Engineering: AWS, Azure Databricks (Spark), ETL Pipelines, FAISS, Git, Docker
Visualization & Business Intelligence: Power BI, Tableau, Matplotlib, Seaborn

EDUCATION

University of New Haven - Tagliatela College of Engineering — **West Haven, CT, USA**
M.S. in Data Science; GPA: 3.89 — Aug 2024 - Expected May 2026

New Horizon College of Engineering — **Bengaluru, India**
B.E. in Computer Science and Engineering; GPA: 3.7 — Jul 2018 - Jul 2022

PROFESSIONAL EXPERIENCE

AI Software Engineer Intern| AI Trusted Advisor | Las Vegas, USA (Remote) — Feb 2026 - Present

- Engineered a GPT-4 based content automation pipeline using prompt engineering and retrieval augmentation, generating 5+ high-quality articles/week and improving organic traffic by 40%.
- Spearheaded website-wide AEO/SEO enhancements by implementing structured data (FAQ, Article schema) and optimizing content for AI search engines, increasing indexed pages by 30% and improving ranking visibility.

AI Engineer| iAssist Innovations Labs (AI Startup) | Bengaluru, India — Jun 2023 - Jul 2024

- Engineered an end-to-end automated ML pipeline for medical bill analysis using Python & scikit-learn, processing 1000+ monthly claims and improving operational efficiency by 25%.
- Reduced manual QA intervention by 65% via root-cause analysis and model optimization, implementing production monitoring to ensure 99%+ uptime and resolve 50+ high-priority issues with sub-5 min MTTR.

Programming Analyst Intern | Cognizant | Bengaluru, India — Sep 2022 - Jan 2023

- Developed and tested full-stack applications using Java, Python, React, and SQL, simulating enterprise-level applications and improving system reliability through debugging and validation.
- Performed data validation and error analysis on application outputs to identify inconsistencies, debug logic, and ensure accuracy and reliability across training modules.

RECENT PROJECTS

ImmigRAG-USA: Hybrid RAG for U.S. Immigration — Oct 2025 - Nov 2025

- Built a production-grade hybrid RAG combining FAISS vector search and BM25 sparse retrieval over 500+ pages of USCIS legal documentation, boosting answer relevance by 30% and cutting LLM hallucination rate from 40% to 12%.
- Fine-tuned and benchmarked 3 open-source LLMs (Llama-3, Mistral, Phi-3) on domain-specific immigration queries.
- Developed and deployed a Dockerized full-stack inference system (FastAPI, Streamlit) with ROUGE-L evaluation and human-in-the-loop validation across 200+ test queries, deployed for scalable inference.

Waste Segmentation System Using Transfer Learning with Mask R-CNN — Sep 2025 - Oct 2025

- Adapted a Mask R-CNN (ResNet-50-FPN) model via transfer learning for multi-class waste instance segmentation, reducing validation loss from 4.40 to 0.86 and achieving 85% classification accuracy.
- Engineered a custom augmentation pipeline (color jitter, Gaussian blur) and applied Grid Search to tune learning rate and weight decay, improving model generalization on unseen waste categories.
- Deployed a real-time inference application on HuggingFace Spaces for automated segmentation of recycling waste.

Real-Time and Batch ETL Pipeline (AWS) — Apr 2025 - May 2025

- Architected a dual-mode AWS e-commerce data pipeline via Kinesis and Glue, ingesting daily product price updates, reducing end-to-end data latency by 60% (from 15 to 6 min).
- Implemented S3 partitioning and Parquet columnar storage, achieving a 3x acceleration in Athena query performance for downstream ML model training and feature engineering.

CERTIFICATIONS

HuggingFace & Anthropic: MCP| DataCamp: Data Scientist Associate | Azure Spark Databricks | PowerBI: Integrating AI

Portfolio: portfolio-mauve-psi-14.vercel.app (live RAG demo) GitHub: github.com/stutibimali (10+ ML projects, active contributions)