Imon Bera

Phone: +1 267-466-7408 | **Email:** ib385@drexel.edu
LinkedIn: linkedin.com/in/imonbera13/ | **GitHub:** github.com/invcble

Professional Summary

AI/ML Specialist with hands-on experience in Retrieval-Augmented Generation (RAG) systems and LLM optimization. Proven expertise in designing AI workflows with Hugging Face Transformers, FAISS vector search, and cloud-native solutions (GCP/AWS). Skilled in Python, TensorFlow, PyTorch, and MLOps tools. Master's candidate in Computer Science (AI/ML) with multiple production-grade implementations of RAG architectures.

Education

Drexel University
Master of Science in Computer Science (GPA: 3.85) | Sep 2023–Present
Specialization: Artificial Intelligence & Machine Learning

Bengal Institute of Technology
Bachelor of Technology in Computer Science (CGPA: 9.14) | Jul 2019–Jun 2023

Technical Skills

Languages: Python, SQL, Java, Scala
ML Frameworks: Hugging Face Transformers, LangChain, TensorFlow, PyTorch, Scikit-learn
Vector Search: FAISS, Qdrant, Weaviate
Cloud Platforms: GCP (Vertex AI, Cloud Run), AWS (EC2), Azure
MLOps Tools: Docker, Kubernetes, Terraform, Pulumi, CI/CD Pipelines
Data Tools: Apache Spark, PyMongo, pandas, NumPy

Professional Experience

AI Engineer Intern | Tapistro, Inc. (San Francisco, CA) | Jun 2024-Dec 2024

- Architected LLM Agentic framework using Retrieval-Augmented Generation (RAG) with FAISS vector search, improving information retrieval accuracy by 35%
- Optimized GPT-4o alternative using quantized Hugging Face models (GGUF/INT8) through vLLM, achieving 90% cost reduction vs commercial APIs

- Designed document ingestion pipelines with Firestore and Vertex AI, processing 10K+ documents daily for RAG systems
- Implemented CI/CD pipelines using GCP Cloud Run and Artifact Registry, reducing deployment times by 40%

Research Assistant | LeBow College of Business (Philadelphia, PA) | Jan 2024-Mar 2024

- Migrated legacy Perl retrieval systems to Python-based solutions with 70% performance improvement
- Developed automated data preprocessing pipelines using pyautogui and Tkinter, reducing manual effort by 80%

Key Projects

Voc-Notes: AI-Powered Lecture Transcription System | Drexel University | Apr 2024-Jun 2024

- Built end-to-end RAG system using Llama3-70b and Mixtral-8x7b LLMs with FAISS vector indexing
- Implemented real-time speech-to-text pipeline with Google Cloud Speech API and Apache Spark preprocessing
- Designed MongoDB document store with PyMongo for efficient retrieval of educational content

Climate Risk Analysis Platform (Chubb Challenge Winner) | Philly Codefest | Apr 2024

- Developed AWS-hosted RAG solution combining statistical models with autoencoder-based damage prediction
- Created Flask web interface with interactive visualization of risk data using JavaScript and Mapbox

Alzheimer’s Classification System | Drexel University | Nov 2023-Dec 2023

- Achieved 86% accuracy in multi-class classification through ensemble of custom Logistic Regression/LDA models
- Implemented image preprocessing pipeline with OpenCV and Canny edge detection

Certifications

- **Reinforcement Learning** (Google/Coursera) | Mar 2023
- **Google Data Analytics Professional Certification** | Aug 2022

Activities

- **Course Assistant:** College of Computing & Informatics | Apr 2024-Present
- **CCI Dean’s Fellowship Recipient** | Jul 2023-Present