HENA KHARWA

📞 +1 732-800-8044 ✉ henakharwa@gmail.com in Linkedin ⦿ Github ⦿ Portfolio

Education

Stevens Institute of Technology   **Sep 2024 – May 2026**
Masters of Science in Applied Artificial Intelligence — GPA - 4.0/4.0 (Through Semester 3)   *Hoboken, New Jersey, USA*
Coursework: Applied Machine Learning, Deep Learning, Modelling and Optimization, Generative AI, Pattern Recognition

Charotar University of Science and Technology   **Oct 2020 – May 2024**
Bachelors of Technology in Computer Engineering — GPA - 3.84/4.0   *Anand, Gujarat, India*
Coursework: Data Structures, Artificial Intelligence, Big Data Analytics, Computer Vision, Cloud Computing

Experience

Stevens Institute of Technology   **April 2025 – Present**
Research Assistant   *Hoboken, New Jersey, United States*
- Built an AI-driven detection system using RAG and embeddings to identify anomalous content with 96% accuracy.
- Fine-tuned LLMs (Grok AI, Hugging Face) with LoRA tuning and prompt optimization, reducing hallucinations by 27%
- Architected scalable AWS ingestion & inference pipelines (EC2, Lambda, S3) for high-throughput ML workloads
- Integrated KGs with FAISS/Pinecone + Flask–Docker APIs for enterprise compliance, raising accuracy by 22%

Dhyey Technologies - TechXi   **Jan 2024 – June 2024**
AI/ML Engineer Intern   *Vadodara, Gujarat, India*
- Designed GPU-optimized deep neural nets in TensorFlow, reduced training time 30% & increased accuracy by 40%
- Used transformers (BERT/GPT) for summarization, embeddings, and fine-tuning, improving relevance by 25%
- Optimized batch & real-time ML inference systems using FastAPI, and Azure Containers with low-latency performance
- Automated ML pipelines with CI/CD (GitHub Actions, Azure ML) and ONNX, reducing deployment time by 30%

Charotar University of Science and Technology   **May 2023 – July 2023**
AI/ML Intern   *Anand, Gujarat, India*
- Modeled hybrid recommender with TensorFlow/PyTorch + CUDA, reducing misclassification 25% & speeding training
- Revamped e-commerce recs with Word2Vec/BERT embeddings, boosting engagement 20% and conversions 5%
- Applied BERT/GPT-2/DistilBERT with FAISS for product suggestions and descriptions, improving accuracy 18%
- Deployed recs services with Docker/FastAPI/Azure ML for scalable real-time inference, reducing latency 25%

Enlighten Infosystems   **May 2022 – July 2022**
Python and ML Intern   *Vadodara, Gujarat, India*
- Established expert system in PyTorch with CUDA, reducing decision time 30% & boosting performance 10%
- Applied OpenCV + deep CNNs for image analysis, improving accuracy 15% and reducing processing time 25%
- Designed Tkinter/Streamlit GUIs with 90% user satisfaction, integrating real-time ML predictions
- Prototyped LLMs (GPT-2/BERT) for summarization/reporting, boosting efficiency 20% & accuracy 15%

Projects

Hospitalist Companion App | *1st Place - Rutgers Health Hack 2025 (RWJUH Spezi App Track)*   **Oct 2025**
- Delivered an intelligent iOS app using SwiftUI + Stanford Spezi Framework, enhancing physician query speed by 2×.
- Integrated GPT-5 + AWS Bedrock RAG pipeline with FHIR-secure backend and vector search, boosting accuracy 35%.
- Trained voice-enabled Conversational AI via AssemblyAI/Firebase-AWS hybrid cloud, supporting real-time responses

AI-Driven Personalized Alumni Engagement System | *Python, Quart, Azure AI Foundry, OpenAI, React* **April 2025**
- Developed Python/Quart + Node.js backend with Vector DBs/embeddings, with match precision 25% on 50K profiles
- Applied LangChain + Azure AI Foundry for LLM orchestration, enabling multi-turn alumni-student recommendations
- Integrated React web app on Azure serverless with Docker + GitHub CI/CD, supporting real-time engagement

PrepRight: AI Powered Interview Preparation App | *Python, Prompt Engineering, LLMs*   **Jan 2025**
- Created dashboards/resume parsers/feedback modules with React.js + Flask APIs with interview assessments on AWS
- Fine-tuned BERT/GPT models via Hugging Face boosting accuracy 25% and deploying low-latency inference APIs.
- Vectorized 1K+ resumes and JDs using NLP embeddings (FAISS) for semantic matching and candidate-job alignment.

Technical Skills

Languages: Python, C++, Java, SQL, Shell Scripting, HTML, CSS, JavaScript
ML/DL: PyTorch/TensorFlow, Scikit-learn, Keras, JAX, OpenCV, MLflow, Matplotlib, Seaborn, Tableau/Power BI
Cloud & Databases: AWS (EC2, S3, Lambda, SageMaker), Azure AI Foundry, Azure Serverless, GCP, PostgreSQL
DevOps & Monitoring: Docker, Kubernetes, FastAPI, Flask, Django, GitHub, CUDA, cuDNN, CuBLAS
Gen AI: LangChain, LlamaIndex, Ollama, BERT, RoBERTa, GPT-2/3.5, FinBERT, PEFT, Prompt Engineering

Certifications

Introduction to Generative AI, AI for Everyone, AI for India 2.0, Artificial Intelligence Virtual Experience Program, Machine Learning, Machine Learning and Big Data Fundamentals, and Python Programming.