Azaan Barlas

(909) 610-0066 | azaanbarlas.com | azaanbarlas@gmail.com | linkedin.com/in/a-barlas | github.com/kingazaan

Education

University of Illinois, Urbana-Champaign Champaign, IL
Master's Degree, Computer Science; Graduate Teaching Assistant for CS 437 *Graduation: May 2024*
Courses: NLP, Machine Learning, Deep Learning, Computer Vision, Cloud Computing, Data Structures, Algorithms, Statistics

Experience

Semiosis AI | Software Engineer *October 2024 – Present*
- Developed a full-stack RAG-based LLM application for FAIR Latter-Day Saints, utilizing embeddings and vector databases to enable semantic search and integrate FAIR information and links into responses, saving search times by 60%
- Designed a multi-agent music management system using CrewAI, Flask, and Cloudflare LLM APIs, creating sub-agent workflows for automating tasks such as event scheduling; boosted engagement by 40% with ChatGPT-like frontend
- Developed a HIPAA-compliant RAG-based healthcare AI chatbot using cloudflare APIs, integrating OpenAI embeddings for medical claims analysis and patient risk scoring. Reduced response latency by 40% while ensuring accurate entity extraction for healthcare policy compliance
- Created spine visualization and pressure analysis training software platform with PyQT GUI for chiropractic clinic; integrated LLM with data analysis capabilities to train chiropractic students on adjustments on Arduino Spine model

School of Information Sciences at UIUC | AI/ML Graduate Researcher *August 2024 – Present*
- Developed a generative AI application: an LLM-based multi-label classifier aimed at predicting labels from 20,000+ research articles using LLaMA 3; achieved 0.83 F1 score after fine-tuning with Low-Rank Adaptation (LoRA) tuning
- Rivaled accuracy and precision of industry benchmark classifier (PubMedBERT) by prompt engineering and Multi-Shot learning on LLaMA instruct models to create cutting-edge multi-label classifier for healthcare research papers

University of Illinois System | AI/ML Software Engineer *June 2023 – July 2024*
- Engineered and deployed an LLM featuring semantic search capabilities using transformers across 1,600 UIUC policy documents for internal use regarding university policy information, reducing search times by $\sim 85\%$
- Built LSTM model using PyTorch to predict bill approval likelihood with an accuracy of 78%; created alert system of highly likely-to-pass bills for administrative affairs team to identify emerging policy trends
- Doubled capacity to research legislative affairs by fine-tuning BERT LLMs to perform aspect-based sentiment analysis on legislative texts, identifying the impact of acts on targeted beneficiaries and affected groups
- Increased college student retention rates by 12% for UIC by using random forest + ARIMA ensemble model to predict continuation and graduation rates per major; identified at-risk students and majors with 83% precision

Esperanto Technologies | Software Engineering Intern *June 2022 – September 2022*
- Implemented Meta's Recommendation model on Esperanto's ET-SoC-1 (microchip) to demo Machine Learning capabilities for clients; hands-on work with large datasets and deploying ML models in production
- Set up an AWS Aurora database, using shell script for data ingestion from S3 bucket containing 1TB data
- Developed an optimized C script to compute summary statistics on a 1TB dataset in under 2 hours, leveraging advanced optimization techniques for performance and scalability

Projects

AI Engineer Project – Healthcare Chatbot (Github)
- Developed a full-stack Django application for a healthcare chatbot, integrating LangChain and enabling AI-driven patient interactions for inquiries, medication management, and appointment scheduling using LLM-agnostic design
- Implemented a Neo4j knowledge graph and PostgreSQL database to store patient data extracted via BioClinicalBERT and regex matching; trained a binary classifier with MedQuAD and Natural Questions datasets

Google Chrome Extension for Entity Recognition using GLINER (Github)
- Developed a Chrome extension utilizing GLINER (Graph-based Lightweight Named Entity Recognition) to extract and highlight entities (names, dates, organizations) from web pages, integrating Python backend with JavaScript frontend
- Optimized the GLINER model for efficient web data processing, ensuring minimal latency and accurate entity extraction across diverse content; demonstrated expertise in NLP, NER, and practical applications of cutting-edge NLP research

HackIllinois John Deere Autonomous Litter Detecting & Mapping Rover (DevPost)
- Programmed a Raspberry Pi rover with motors and camera using Python and OpenCV for object detection; fine-tuned a lightweight YOLO model with the TACO dataset (10,000 images) for autonomous traversal and litter detection
- Integrated hardware and software components for real-time processing and mapping; finalist in the Hack Knights path, showcasing skills in embedded systems, computer vision, and robotics

Technical Skills

Languages and Frameworks: Python, JavaScript, TypeScript, C++, SQL, R, Raspberry Pi, Arduino IDE, HTML, CSS, Django, Flask, Node.js, React.js, TensorFlow, PyTorch, Scikit-learn, Hugging Face Transformers, NumPy, Pandas
Developer Tools: AWS (S3, EC2, SageMaker, Lambda, RDS, CLI), GCP (BigQuery, Vertex AI, Cloud Functions), Azure, Cloudflare, Docker, Kubernetes, Git, GitLab, Linux/Unix, PostgreSQL
Concepts: Machine Learning Recommendation Systems, Large Language Models (GPT-4, BERT, Llama), Retrieval-Augmented Generation (RAG), Cloud ML Deployment, ETL Pipelines, Data Engineering, Web Scraping