Bargav Jagatha

857-398-8179 | jbargav025@gmail.com | linkedin.com/in/bargav/ | github.com/bargav25 | Boston, MA

Experience

Research Scientist May 2024 – Present
Chobanian & Avedisian School of Medicine, Boston University *Boston, MA*

- Engineered **attention-based imputation models** for Alzheimer's patient data on longitudinal cognitive assessments (GERAS and HRS), achieving **0.43 MAE** (US) and **0.44 MAE** (EU) in GERAS, **15%** improvement over previous methods.
- Further evaluated imputation models using several downstream tasks like **forecasting** and statistical analyses.
- Developed and deployed a clinical **RAG-based chatbot** using **DeepSeek-R1** and **vLLM**, integrating with **ClinicalTrials.gov** data to provide real-time information access.

Data Scientist (Machine Learning Engineer) Feb. 2022 – July 2023
OLA, ANI Technologies *Bangalore, India*

- Led end-to-end development and **production deployment** of **dynamic pricing and rider-driver matching ML models** for OLA Cabs, driving a **6% GMV** increase across the platform.
- Built comprehensive **MLOps infrastructure** for these models, including **automated retraining pipelines**, **drift detection systems**, and **performance monitoring dashboards** to ensure production reliability.
- Implemented rigorous **shadow analysis** and **A/B testing** methodologies, significantly reducing deployment failures across cities throughout India.
- Optimized model inference latency by **70%** through efficient code vectorization and hardware acceleration, enabling real-time pricing decisions during peak demand periods.
- Architected and deployed an **LSTM-based proximity unlock system** (**97% accuracy**) for the **Android HMI** of eBikes, integrating ML models with embedded systems.
- Developed a **Graph Neural Network** ETA prediction model achieving a **1.3-minute MAE** overall, with significant improvements for short-distance driver arrival estimates.
- Deployed a **customer support chatbot** (**96% intent accuracy**) using **GitLab CI/CD**, **Kubernetes**, and **Vue.js**.

Edge AI Intern May 2021 – July 2021
Samsung Research *Bangalore, India*

- Designed an SQLite database to store clip metadata and implemented several Java functions (Android interface) for efficient management and storage of smart home security clips on edge devices.
- Developed **performance-optimized native C code** using **FFmpeg** to convert H264-encoded videos to JPEG/PNG images for thumbnail extraction.

Education

Boston University Boston, MA
*Master of Science in **Artificial Intelligence - CGPA 3.9*** *Sep. 2023 – Jan. 2025*
***Head Teaching Assistant** (2x) for Graduate-level **Data Engineering at Scale** course*
Led weekly coding discussions covering ETL, Docker, Spark, Hadoop, Kubernetes, AirFlow, Kafka
Coursework: Tools for Data Science, NLP, Reinforcement Learning, Principles of ML, CV, Deep Learning

National Institute of Technology Bhopal, India
Bachelor of Technology in Computer Science and Engineering *July 2018 – May 2022*

Technical Skills

Languages: Python, R, C/C++, SQL, CUDA, Triton

Frameworks: PyTorch, TensorFlow, scikit-learn, JAX, PySpark, Hugging Face, vLLM

Cloud/MLOps: AWS (SageMaker, Lambda), GCP, Docker, Kubernetes, CI/CD

LLM/Big Data: RAG, LLM Inference Optimization, PEFT, LoRA, RLHF, Hadoop, Spark, Kafka

Projects

eBike Finder | *Python, AWS Lambda, S3, EC2, XGBoost, SQLite3, Gradio* Sept. 2024

- Engineered a **production-ready ML service** for predicting e-bike availability across Boston's Bluebikes stations with **90% accuracy**, deployed on Hugging Face Spaces.
- Architected a complete **AWS infrastructure** with Lambda functions for data collection, S3 for storage, EC2 for model hosting, and API Gateway for frontend integration.
- Implemented an **XGBoost model** with automated retraining pipelines to incorporate real-time Bluebikes GBFS feed data, improving prediction accuracy over time.

Dynamic NeRF for Real-Time 3D Scene Reconstruction | *PyTorch, CUDA, NeRF* Sept. 2024 – Nov. 2024

- Developed a novel **keypoint-based NeRF architecture** that eliminates the need for traditional SfM/COLMAP.
- Implemented a **custom view synthesis algorithm** allowing users to generate new perspectives and animations from a single input video.

Intelligent Grammar Correction & Paraphrasing Bot | *Python, HuggingFace, DPO, KTO, RLHF* Aug. 2024

- Implemented **SmolLM architecture from scratch** and fine-tuned it using a sequential training strategy: SFT followed by **reinforcement learning**.
- Applied **DPO** (Direct Preference Optimization) and **KTO** (KL-constrained Threshold Optimization) techniques to align the model with human preferences.
- Demonstrated clear improvement in grammar correction quality through rigorous **A/B testing** against baseline models.

Multimodal RAG System for Open-Domain Retrieval | *Python, Qdrant, ColPali* Feb. 2025

- Engineered a **high-performance vector database** using Qdrant to index hundreds of GBs of multimodal Wikipedia data, including text and images.
- Orchestrated batched inference on multiple **A100 GPUs** to improve the speed of generating embeddings and querying from the database.
- Implemented a **multimodal RAG pipeline** capable of generating text based on both textual and visual information from retrieved documents.

Vehicle Classification System for Supply Chain Logistics | *AWS SageMaker, S3, EC2, Lambda, CloudWatch* Feb. 2023

- Developed an end-to-end **vehicle classification model** distinguishing between cars, bikes, and other vehicles, deployed as a fully managed **SageMaker inference endpoint**.
- Architected a complete **AWS ML infrastructure** with **S3** for data storage, **EC2** for preprocessing, **Lambda functions** for event-driven triggers, and **SageMaker pipelines** for training.
- Implemented comprehensive **monitoring and observability** using **CloudWatch** metrics, alarms, and automated alerts to ensure model reliability in production.
- Selected for the **AWS AI/ML Advanced Scholarship Program** by clearing the AWS Deep Racer League, earning a **Gold Badge** for successfully implementing this end-to-end vehicle classification system.

Certificates & Achievements

- **Silver Medal** in the **AI Math Olympiad** on Kaggle: Fine-tuned **Gemma** and using **CoT + Self-Consistency** techniques on the **DeepSeekMath** model. Currently competing in the second version using **Qwen1.5B** and training with **GRPO** (Gradient Reward in Preference Optimization) while merging reward signals from correctness and PRM reward models.
- Won the **iNeuron AI/ML Hackathon** by designing an advanced **RASA chatbot**, securing first place and a $2,500 prize.

Publications

- **Imputation of Missing Cognitive Assessment Scores in Alzheimer's Disease: A Self-Attention Based Deep Learning Approach** (In Progress)
- **Solving the International Mathematical Olympiad, Harvard University Mathematics PhD Qualifying Exams, and MIT's EECS Curriculum at a Human Level** (In Progress)