Harsh Lara

412-583-1116 | hlara@alumni.cmu.edu | harshlara.github.io | harshlara | harshlara

Work Experience

Google DeepMind — *Mountain View, CA*
Software Engineer — *Sept 2021 - Present*
- Lead for fine-tuning Project Astra for tool-use
- Worked on 'Improve Mathematical Reasoning in Language Models by Automated Process Supervision' in H2 of 2024
- Worked on 'MoDE: Effective Multi-task Parameter Efficient Fine-Tuning with a Mixture of Dyadic Experts' in H1 of 2024
- Presented the 'Evaluation of Synthetic Datasets for Conversational Recommender Systems' paper at NeurIPS 2022
- Filed multiple international patents for technology using LLMs
- Implemented critical end-to-end features, including LoRA adapters for internal models

Microsoft — *Bellevue, WA*
Data & Applied Scientist, Azure Machine Learning — *Dec 2019 - Sept 2021*
- Implemented the end-to-end infrastructure for training and inference using neural-networks for text data-labelling
- Implemented a model wrapper that resulted in the reduction of AutoML-shipped model size by 95%
- Led a ML Pipelines infrastructure transfer project that required significant cross-team collaboration

Apple — *Cupertino, CA*
Machine Learning Engineer Intern, Siri
- Developed an API for dynamic interpolation of probabilistic language-models on-device
- Achieved a multi-domain increase in speech-to-text (recognition) accuracy
- Presented a live demo of the feature to AI/ML senior leadership

Education

Carnegie Mellon University, School of Computer Science — *Pittsburgh, PA*
MS, Computational Data Science
- Selected courses: Cloud Computing | Computer Systems | Intro to Machine Learning | Language and Statistics | Large-Scale Multimedia Analysis | Applied Machine Learning | Deep Learning | Computational Ethics for NLP

National Institute of Technology Karnataka (NITK) — *Surathkal, India*
B Tech, Computer Science and Engineering
- Selected courses: Operating Systems | Databases | Distributed Systems | Object Oriented Programming | Advanced Algorithms | Data Structures | Image Processing | Data Communication | Graph Theory | Discrete Math | Linear Algebra

Skills

Strengths: Artificial Intelligence, Language Technologies, Machine Learning, Deep Learning, Cloud Computing
Languages/Libraries: Python (Advanced), C/C++ (Advanced), Java (Intermediate), SQL (Intermediate)
Tools/Technologies: PyTorch, Azure, AWS, GCP, Kafka, Samza, Apache Spark, JUnit, Git, Perforce, LaTeX

Projects

Evaluation of Synthetic Datasets for Conversational Recommender Systems
- Developed a framework for evaluating synthetic datasets generated using Large-Language Models (LLMs)
- Presented my work at the Human Evaluation of Generative Models (HEGM) Workshop hosted at NeurIPS 2022

Persistent Storage of Dialogue Data for the Next-Gen Recommender System Demo
- Implemented an end-to-end storage solution for collecting conversational AI data while preserving user-privacy
- Designed and implemented differential access to storage, in order to protect sensitive dialogue data

Machine Learning Assisted Text-Labeling on the Cloud
- Implemented text-labeling assistant using ML classifiers
- Leveraged open-sourced BERT-based models in order to build the infrastructure for the data labeling solution

Semantic Interpretation of Aggression from Code-mixed Social Media Posts
- Implemented a classifier using hierarchical attention networks to detect aggression in code-mixed Twitter posts
- Classified tweets into a wide aggression typology - overt, covert and non-aggressive
- Used word-level embedding to achieve an accuracy of 74 %

Horizontal Scaling and Advanced Resource Scaling on AWS
- Designed a VM network and invoked cloud APIs programmatically to provision cloud resources for a dynamic load
- Configured and deployed an Elastic Load Balancer (ELB) along with an Auto Scaling Group on AWS for backend infrastructure
- Developed fault-resistant elasticity policies to maintain QoS of web service

Publications

- MoDE: Effective Multi-task Parameter Efficient Fine-Tuning with a Mixture of Dyadic Experts, NAACL 2025
- Improve Mathematical Reasoning in Language Models by Automated Process Supervision, arXiv 2024
- Evaluation of Synthetic Datasets for Conversational Recommender Systems,
  Human Evaluation of Generative Models (HEGM) Workshop hosted at NeurIPS 2022
- Leveraging Large Language Models in Conversational Recommender Systems, arXiv 2023