Harsh Lara

412-583-1116 | hlara.tech@gmail.com | harshlara.github.io | harshlara | harshlara

Work Experience

| Google DeepMind — *Mountain View, CA*

Software Engineer — *Sept 2021 - Present*

- Fine-tuning lead for tool-use in *Project Astra* for productivity (supporting Calendar, Contacts, Gmail & Keep)
- Worked on *Improve Mathematical Reasoning in Language Models by Automated Process Supervision*
- Worked on *MoDE: Effective Multi-task Parameter Efficient Fine-Tuning with a Mixture of Dyadic Experts*
- Presented the *Evaluation of Synthetic Datasets for Conversational Recommender Systems* paper at NeurIPS 2022
- Filed multiple international patents for technologies leveraging 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%

| Apple — *Cupertino, CA*

Machine Learning Engineer Intern, Siri — *Summer 2019*

- 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

Effective Multi-task Parameter Efficient Fine-Tuning of LLMs — *Google DeepMind*

- Developed a parameter-efficient framework for multi-task adaptation in LLMs
- Designed a novel task-specific adapter architecture that outperforms existing methods in multi-task fine-tuning
- Demonstrated significant performance improvements in handling over 700 tasks while maintaining computational efficiency
- Contributed to advancing lightweight, high-performance models for a wide range of applications

Improve Mathematical Reasoning in LLMs by Automated Process Supervision — *Google DeepMind*

- Developed OmegaPRM, an efficient MCTS algorithm that automates the collection of high-quality process supervision data
- Engineered a divide-and-conquer approach within OmegaPRM to swiftly identify errors in Chain-of-Thought reasoning,
- Collected over 1.5 million process supervision annotations, training Process Reward Models, significantly boosting LLM performance on complex mathematical reasoning
- Improved LLM performance from 51% to 69.4% on MATH500 and from 86.4% to 93.6% on GSM8K

Evaluation of Synthetic Datasets for Conversational Recommender Systems — *Google Research*

- Led the development of a framework for evaluating synthetic datasets generated using Large-Language Models (LLMs)
- Presented work at the Human Evaluation of Generative Models (HEGM) Workshop hosted at NeurIPS 2022

Leveraging LLMs in Conversational Recommender Systems — *Google Research*

- Developed RecLLM, a conversational recommender system for YouTube videos utilizing LaMDA
- Designed an architecture integrating LLMs to improve dialogue management, and provide explainable recommendations
- Developed a controllable LLM-based user simulator to generate synthetic data, addressing data scarcity challenges
- Demonstrated RecLLM's fluency and diverse functionalities through illustrative conversations, showcasing its potential in personalized recommendation systems

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