Maryam Haghifam

Email LinkedIn GitHub Google Scholar Website 2133125448 Los Angeles, CA, USA

Research Interests

- Large Language Model, Efficient Deep Learning, Speculative Decoding, Long-Context LLMs

Education

- **University of California, Los Angeles (UCLA)** Los Angeles, USA
  PhD in Computer Science *January 2025 – December 2029*
  Supervisor: Prof. Yizhou Sun
  - Courses: Large Scale Social and Complex Networks, Reinforcement Learning
- **University of Toronto** Toronto, Canada
  MSc in Computer Science *September 2021 – August 2023*
  Supervisors: Prof. Gennady Pekhimenko and Prof. Nandita Vijaykumar
  - Courses: Neural Networks and Deep Learning, Computer Graphics, Applications in Parallel Programming
- **University of Tehran** Tehran, Iran
  BSc in Electrical Engineering (Area: Control Theory) *September 2016 – June 2021*
  Minor in Computer Engineering
  Supervisor: Prof. Vahid Shah-Mansouri
  - Courses: Linear Algebra, Statistical Inference, Advanced Programming, Artificial Intelligence, Algorithm Design, Data Structures, Probability and Statistics

Work Experience

- **Machine Learning Researcher** Robust and Interpretable Machine Learning Lab
  Supervisor: Prof. Mohammad Hossein Rohban *July 2024 - December 2024*
  - Worked on the robustness of anomaly detectors guided by Vision-Language Models.
  - Used Vision-Language Models to significantly improve model accuracy.
- **Computer Vision and AI Research Engineer** EAIGLE Inc.
  R&D Team *June 2023 - March 2024*
  - Enhanced customer safety at Walmart stores, via detecting liquid spillages and blockages in emergency exit doors, through object detection (YOLOv6) and segmentation (U-Net).
  - Demonstrated a remarkable accuracy of **98%** within surveillance videos from Walmart stores.
  - Designed a *no-code/low-code* framework to enable non-expert individuals to train and deploy deep learning models.
  - Reduced the cost of annotating segmentation masks by using cheaper point annotations to fine-tune Meta's *Segment Anything Model* in a weakly supervised manner.
  - Enhanced the accuracy of *person re-identification* methods by processing local features at body key points using Graph Convolutional Networks.
  - Designed a system for camera placements and calibration, leveraging Neural Radiance Fields and an attention mechanism.

Research Experience

- **Graduate Research Assistant** University of California, Los Angeles
  Supervisors: Prof. Yizhou Sun and Prof. Jason Cong *January 2025 – Present*
  - **Octopus: Multi-expert self-speculative decoding** Designed a multi-expert framework with shallow, few-layer MLP draft heads and multi-token training.
  - Achieved 80% acceptance rate and 2× end-to-end speedup; draft-head entropy comparable to transformer-based draft heads; implemented the full training approach; reduced draft–target communication during training and inference.

- **Long-context QA over Siemens technical manuals** Generated QA pairs from HTML/XML handbook files and prepared data for training.
- Applied Hierarchical Memory Transformer for long-context QA and used the semantic hierarchy of the dataset.
- Delivered ≈20% ROUGE improvement over the baseline.

- **Graduate Research Assistant** University of Toronto
  Supervisors: Prof. Gennady Pekhimenko and Prof. Nandita Vijaykumar *September 2021 - August 2023*
  - Developed a multi-stage framework to preserve architecture privacy in runtime optimization of large models.
  - Adopted an off-the-shelf graph partitioning algorithm to divide the computation graph into smaller subgraphs.
  - Used Graph RNN and an SMT solver to generate similar replicas of the subgraphs, thus obfuscating them.
  - Demonstrated negligible harm to runtime optimization caused by our privacy-preserving method.

- **Undergraduate Research Assistant** University of Tehran
  Supervisor: Prof. Vahid Shah-Mansouri *June 2020 - December 2020*
  - Analysed various methods of predicting network traffic patterns for enhancing the quality of service.
  - Proposed a novel approach based on reinforcement learning to predict network traffic patterns, outperforming the conventional ARMA, ARIMA, and RNNs.

- **Undergraduate Research Assistant** University of Toronto
  Supervisor: Prof. Yashar Ganjali *Septermber 2020–November 2020*
  - Evaluated predictability of network flows for resource provisioning based on reinforcement learning, demonstrating high correlation within network traffic data.

- **Undergraduate Research Intern** University of Tehran
  Supervisor: Prof. Vahid Shah-Mansouri *April 2019–June 2019*
  - Implemented a crowdsourcing platform requiring crowd workers to report their confidence.
  - Evaluated our platform through statistical analysis, demonstrating a decrease in crowdsourcing mistakes.

Technical Skills

- **Programming**: Python (PyTorch, TensorFlow, OpenCV, PyTorch Geometric), C/C++
- **Cloud Services**: Amazon Web Service (AWS)

Publications

- Y. Gao, **M. Haghifam**, R. Tu, C. Giannoula, G. Pekhimenko, N. Vijaykumar, "*Proteus: Preserving Model Confidentiality during Graph Optimizations.*" MLSys 2024.
- S. Nili Ahmadabadi, **M. Haghifam**, V. Shah-Mansouri, S. Ershadmanesh, "*Design and Evaluation of Crowd-sourcing Platforms Based on Users' Confidence Judgments.*" Scientific Reports journal.

Selected Course Projects

- **Distributed Training for Neural ODE**
  - Designed a novel method for distributed training of Neural ODEs.
  - Accelerated prediction of multivariate time-series in Neural ODEs, by inferring the low-frequency variables from the higher-frequency variables.
- **Efficient Transformers with Random Attention**
  - Utilized random sampling of attention matrices in transformers to enhance the performance.
  - Demonstrated faster training while preserving the accuracy of the model.
- **UTrello**
  - Implemented a project management tool, inspired by the *Trello* platform, using object-oriented design in C++.

Honors and Awards

- Received research support from the University of California, Los Angeles.
- Received full graduate fellowship at the University of Toronto
- Ranked top 15% in the Electrical Engineering major out of 120 undergraduate students
- Ranked top 0.3% among 160,000 participants in Iran's Nationwide University Entrance Exam

Teaching Assistantship Experience

- **University of Toronto**: Data Structures and Algorithms, Introduction to Programming
- **University of Tehran**: Computer Networks, Engineering Mathematics, Introduction to Computing Systems and Programming, Principals of Electronics (Lab)

Conference Reviewer Experience

- **MLSys 2023**: Artifact and paper code reviewer
- **EuroSys 2022**: Paper reviewer

Volunteering Experience

- **Clean Code Workshop Conductor**
  Karyar College *January 2022 and July 2023*
  - Presented workshops to unprivileged talented individuals on the best practices for writing clean codes.
- **Python Teaching Assistant**
  Karyar College *May 2021 - September 2021*
  - Guided a group of unprivileged talented individuals in weekly virtual sessions teaching Python.
  - Developed personalized learning plans for each student.
- **Event Organizer**
  Bahar Charity *July 2022 - October 2022*
  - Organized events to raise awareness on education and empowerment of unprivileged groups.