Maryam Haghifam

maryamhgf@cs.ucla.edu · LinkedIn · GitHub · Scholar · Website · 213 312 5448 · Los Angeles, CA

Education

- **University of California, Los Angeles (UCLA)** — Los Angeles, USA
  Ph.D. in Computer Science, Supervisor: Prof. Yizhou Sun — *Jan 2025 – Expected Dec 2029*
  - *Relevant Coursework:* Neural Networks & Deep Learning II, Large-Scale Data Mining, Artificial Intelligence, Reinforcement Learning, LLM Inference on Edge Devices.
- **University of Toronto** — Toronto, Canada
  M.Sc. in Computer Science, Supervisors: Profs. Gennady Pekhimenko, Nandita Vijaykumar — *Sep 2021 – Aug 2023*
- **University of Tehran** — Tehran, Iran
  B.Sc. in Electrical Engineering (Control), Minor in Computer Engineering — *Sep 2016 – Jun 2021*

Research Experience

- **Graduate Research Assistant** — University of California, Los Angeles
  Supervisors: Profs. Yizhou Sun, Jason Cong — *Jan 2025 – Present*
  - **SeDeM: Selective Decompression of Hidden-State Memories** (*EMNLP 2026 Main Conference*). Compresses long contexts into hidden-state memory blocks, retrieves relevant blocks via late-interaction selection, and decompresses them; improves accuracy over compression baselines.
  - **Octopus: Multi-Expert Self-Speculative Decoding.** Shallow MLP draft heads with multi-token training; 2× end-to-end speedup, 80% speculation acceptance rate.
  - **H²MT: Hierarchical Semantic Memory Transformer** (*COLM 2026 Workshop*). Routes retrieval by document structure; cuts TTFT up to 90% and peak memory ∼20% vs. baselines, with strong ROUGE-L on structured QA.
- **Graduate Research Assistant** — University of Toronto
  Supervisors: Profs. G. Pekhimenko, N. Vijaykumar — *Sep 2021 – Aug 2023*
  - **Proteus** (*MLSys 2024*): compute-graph partitioning and Graph RNN/SMT-based subgraph obfuscation to preserve model confidentiality.

Industry Experience

- **AI/ML Research Intern – PhD Level** — Optum Technology, UnitedHealth Group
  Optum Tech AI/ML Division, Remote, CA — *Jun 2026 – Present*
  - Design a temporal personalized-memory architecture for longitudinal healthcare AI agents, separating current from superseded patient facts via active/archive memory stores linked by knowledge-graph relations.
  - Build the memory write/query pipeline – extraction, impact retrieval, state updates, relation linking, retrieval and reranking for grounded responses – and evaluate on a longitudinal patient-memory QA benchmark.
- **Machine Learning Researcher** — Robust and Interpretable Machine Learning Lab
  Supervisor: Prof. Mohammad Hossein Rohban — *Jul 2024 – Dec 2024*
  - Studied robustness of vision-language-guided anomaly detectors under distributional shift.
- **Computer Vision and AI Research Engineer** — EAIGLE Inc.
  Toronto, Canada — *Jun 2023 – Mar 2024*
  - Built detection (YOLOv6) and segmentation (U-Net) pipelines for retail safety monitoring, reaching 98% accuracy; designed a low/no-code training workflow for non-experts and reduced labeling cost via weakly supervised SAM fine-tuning.

Publications

- **M. Haghifam**, J. Cong, Y. Sun, "*SeDeM: Selective Decompression of Hidden-State Memories for Long-Context Question Answering*." **EMNLP 2026 Main Conference**.
- **M. Haghifam**, Z. He, J. Cong, Y. Sun, "*H^2MT: Semantic Hierarchy-Aware Context-Time Memory for Beyond-Window Language Modeling*." **COLM 2026 Workshop** (Context Beyond the Window).
- Y. Gao, **M. Haghifam**, *et al.*, "*Proteus: Preserving Model Confidentiality during Graph Optimizations*." MLSys 2024.
- C. Chang, Y. Xiao, *et al.*, **M. Haghifam**, *et al.*, "*Training Recipes for Agentic Reinforcement Learning in LLMs: A Survey*." TechRxiv preprint, 2026.
- S. Nili Ahmadabadi, **M. Haghifam**, *et al.*, "*Design and Evaluation of Crowdsourcing Platforms Based on Users' Confidence Judgments*." Scientific Reports.

Technical Skills

- **Programming:** Python, C++ | **ML / LLM:** PyTorch, Hugging Face, TensorFlow/Keras, scikit-learn
- **ML Systems:** DeepSpeed, Slurm, ONNX | **Tools:** Git, Docker | **Cloud:** AWS, Azure