

# Malikeh Ehghaghi

LinkedIn: [Link](#) | Twitter: [Link](#) | Google Scholar: [Link](#)

Personal Website: [Link](#)

Toronto, ON, CA

Email: malikehehghaghi@gmail.com

Phone number: +1 (647) 9751375

## Research Interests

---

- **LLM Safety & Security:** Safety evaluation and benchmarking, automated red-teaming, LLM safeguarding, and risk mitigation
- **Efficient & Scalable ML:** Decentralized training, modular AI, model merging, long-context learning, and test-time training
- **Trustworthy AI:** Factuality, interpretability, and fairness

## EDUCATION

---

### Ph.D. in Computer Science

September 2025 - Present

Academic Advisor: Colin Raffel

University of Toronto, Department of Computer Science

### MSc in Applied Computing

September 2020 – December 2021

Academic Advisor: Frank Rudzicz

Industry Advisor: Jekaterina Novikova

University of Toronto, Department of Computer Science

### BSc in Computer Engineering

September 2015 – March 2020

Academic Advisor: Siamak Mohammadi

University of Tehran, Department of Electrical and Computer Engineering

## EXPERIENCE

---

### Vector Institute, Toronto, ON, CA

February 2025 – Present

PhD Student (ML Research Scientist) - Supervisor: Colin Raffel

- **R3 Lab:**
  - Conducting research focused on decentralizing, democratizing, and de-risking large-scale AI.
  - Active Projects:
    - FineSearch: Reducing hallucinations in long-context QA through memory-efficient retrieval-augmented architectures trained on large-scale synthetic data.
    - TokSuite: Introducing a controlled framework for studying the impact of tokenization on language models by training identical models with different tokenizers. Introduces a real-world perturbation benchmark to analyze how tokenization strategies affect performance and robustness systematically.
    - pRisk-Pressure: Introducing a new metric for adversarial robustness that models LLM safety as a continuous risk–pressure curve instead of fixed-budget attack success rates. Varies attack refinement steps to capture attack cost-efficiency, enabling finer-grained comparisons across model families, training phases, and attack strategies.

### Arcee AI, San Francisco, CA, US

January 2024 – February 2025

Applied NLP Research Engineer

- **NLP Research Engineering Team:**

- Engineering & Tooling:
  - Contributed to Arcee's MergeKit (7K GitHub stars), the largest open-source model merging toolkit.
  - Contributed to Arcee's DistilKit (~1K GitHub stars), a toolkit for distilling large teacher models into smaller, efficient ones.
  - Built core components of Arcee's SaaS Platform, an end-to-end pipeline for efficient LLM training and inference.
  - Contributed to developing Arcee Orchestra, an in-house platform for designing agentic workflows.
  - Built multi-GPU training pipelines:
    - On AWS Trainium for domain adaptation of LLMs.
    - On Together AI Cluster using Megatron-LM for training Mixture-of-Experts (MoE) models.
- Project Leadership:

- Contributed to building an automated workflow generation and large action planning for personal assistant bots in collaboration with SKTelecom
- Built an agentic HR assistant bot in partnership with Riviera Partners
- Built function-calling agent models for banking transactions in partnership with American Express (AmEx)
- Implemented RAG-based Las Vegas tourism search system in partnership with MongoDB
- Built KidsSafe LLMs with Angel Kids Inc., safe and personalized models for children aged 5–12
- o Multilingual & Social Impact Initiatives
  - Led the Arcee Globe initiative to build culturally and linguistically specific LLMs
    - Released Arcee Meraj and Meraj Mini, top-performing Arabic LLMs
- o Evangelism & Communication
  - Co-hosted the Small Language Models (SLMs) Show (with Julien Simon), covering SOTA in GenAI

### **Lavita, Palo Alto, CA, US**

August 2023 – December 2023

Machine Learning Consultant

- **Machine Learning Research Team:**

- o Contributed to designing and implementing a framework for a general-purpose decentralized medical chatbot in partnership with Harvard Medical School and Dartmouth Center for Precision Health and Artificial Intelligence.

### **University of Toronto, Toronto, Canada**

January 2023 – July 2023

Machine Learning Research Team Lead - Supervisor: Prof. En-Shiun Annie Lee

- **Lee Lab:**

- o Led a team of graduate and undergraduate students working on the medical interpretability of clinical notes
- o Benchmarked deep learning models to detect disease labels from clinical notes collected from electronic health records (EHRs)
- o Implemented model-agnostic and model-specific NLP interpretability techniques (e.g., SHAP, LIME, probing, adversarial techniques, etc.) to explore the most important input features in predicting the diagnosis labels
- o Benchmarked interpretability methods based on different interpretability evaluation techniques (e.g., human-centric evaluation, comprehensiveness, sensitivity, etc.)

### **Winterlight Labs Inc./Cambridge Cognition Inc., Toronto, Canada**

April 2021 – July 2023

Machine Learning Engineer - Team Lead: Jekaterina Novikova

- **Machine Learning Research Team:**

- o Led client and research projects with a focus on monitoring and detection of cognitive impairment (e.g., Alzheimer's disease) and mental disorders (e.g., depression, Schizophrenia, anxiety, etc.) from speech and language data in multilingual settings in partnership with Johnson and Johnson and Genentech Roche Group.
- o Developed speaker verification tools for clinical trials to detect duplicate raters and participants with transferability to non-English languages (e.g., German, Danish, Spanish, and Arabic)
- o Analyzed the robustness of the disease detection models to different sources of noise, and developed tools to assess the interpretability and fairness of disease detection models for mental health disorders and cognitive impairment

### **KavoshComAsia Company, Tehran, Iran**

February 2019 – February 2020

Machine Learning Research Assistant – Supervisor: Siamak Mohammadi

- **Research and Development Division:**

- o Designed and implemented on-device ECG heartbeat classification models for arrhythmia, applying deep convolutional neural networks and sequence-to-sequence models in the TensorFlow framework

### **York University, Toronto, Canada**

Summer 2019

Machine Learning Research Intern – Supervisor: Prof. Aijun An

- **Data Mining Lab:**

- o Developed an unsupervised recommender system tool based on the shopping behavior and lifestyle of a group of people in collaboration with Manifold Data Mining Inc.

### **University of Tehran, Tehran, Iran**

July 2018 – February 2019

Machine Learning Research Assistant – Supervisor: Prof. Mehdi Tale Masouleh

- **Human and Robot Interaction Laboratory:**

- o Implemented a motion planning and indoor localization system for Sanbot Robot, applying machine learning algorithms, including reinforcement learning, Particle Swarm Optimization (PSO), etc., in the TensorFlow framework

## PEER-REVIEWED PUBLICATIONS

---

Goddard, C., Siriwardhana, S., **Ehghaghi, M.**, Meyers, L., Karpukhin, K., Benedict, B., McQuade, M., Solawetz, J., “Arcee's MergeKit: A Toolkit for Merging Large Language Models” - EMNLP 2024 Industry Track - Accepted

**Ehghaghi, M.**, Zhou, P., Rajabi, S., Cheng, W., Kuo, C., Lee, E., “Interpretable Disease Prediction from Clinical Text by Leveraging Pattern Disentanglement”, IEEE BHI 2023 - Accepted

**Ehghaghi, M.**, Stanojevic, M., Akram, A., and Novikova, J., “Factors Affecting the Performance of Automated Speaker Verification in Alzheimer's Disease Clinical Trials”, 5th Clinical Natural Language Processing Workshop, ACL 2023 - Accepted

**Ehghaghi, M.**, Rudzicz, F., and Novikova, J., “Data-driven Approach to Differentiating between Depression and Dementia from Noisy Speech and Language Data”, 8th Workshop on Noisy User-generated Text, COLING 2022 - Accepted

Tasnim, M., **Ehghaghi, M.**, Diep, B., and Novikova, J., “Depac: a corpus for depression and anxiety detection from speech”, 8th *Workshop on Computational Linguistics and Clinical Psychology*, NAACL 2022 - Accepted

**Ehghaghi, M.**, Novikova, J., Sett, A., Hejrati, M., Robin, J., Teng, E., and Hashemifar S., “Benchmarking Prognostic Longitudinal Machine Learning Models of Alzheimer's Disease Using Speech Features”, 1st Workshop on Applications of Medical AI, MICCAI 2022 - Accepted

## PREPRINTS

---

Altıntaş, Gül Sena\*, **Ehghaghi, M.\***, Lester, B., Liu, F., Zhao, W., Ciccone, M., Raffel, C., “TokSuite: Measuring the Impact of Tokenizer Choice on Language Model Behavior” - Under Review (ICML 2026)

Akram, A., **Ehghaghi, M.**, Stanojevic, M., and Novikova, J., “Zero-Shot Multilingual Speaker Verification in Clinical Trials” - Available on arxiv.org

Gauthier-Caron, T., Siriwardhana, S., Stein, S., **Ehghaghi, M.**, Goddard, C., McQuade, M., Solawetz, J., Labonne, M., “Merging in a Bottle: Differentiable Adaptive Merging (DAM) and the Path from Averaging to Automation” - Available on arxiv.org

## WORKSHOPS

---

- The First Workshop on Machine Learning for Cognitive and Mental Health (ML4CMH), AAI 2024 - Program Co-Chair

## INVITED TALKS

---

- “Women in Industry Research”, Build With AI: Women in Tech (WIT) conference 2026 by Google Developer Groups (University of Toronto) and ACM-W: ACM's Women in Computing - Panelist
- “Model Merging: Theory, Practice and Applications”, In-person Tutorial, NeurIPS 2025 - Speaker
- “Scaling Down, Powering Up: Can Efficient Training Beat Scaling Laws”, In-person Workshop, Toronto Machine Learning Summit (TMLS) 2025 - Speaker
- Panel Discussions with Female Leaders in the GTA in Tech, Girls in Tech Conference (GITCon) 2025 - Panelist
- “From Bias to Brilliance: How Women Are Rewriting the Rules of AI”, Women in Tech (WIT) Conference 2025 by Google Developer Groups (GDG) and ACM-W: ACM's Women in Computing - Speaker
- Women in AI Research (WiAIR) Podcast, 2025 - Co-host
- “Factors Affecting the Performance of Automated Speaker Verification in Alzheimer's Disease Clinical Trials” paper at 5th Clinical Natural Language Processing Workshop, ACL 2023 - Presenter
- “Data-driven Approach to Differentiating between Depression and Dementia from Noisy Speech and Language Data” paper at 8th Workshop on Noisy User-generated Text, COLING 2022 - Presenter
- Wo(Men) In Tech Program, Google Developer Student Clubs, Fall 2021 - Industry speaker

## TEACHING EXPERIENCE

---

**Agentic AI Bootcamp 1.0 and 2.0, Vector Institute, Toronto, Canada**  
Technical facilitator

Summer 2025 - Winter 2026

**Agentic AI Evaluation Bootcamp 1.0, Vector Institute, Toronto, Canada**  
Technical facilitator

Winter 2026

**University of Toronto in Partnership with Vector Institute, Toronto, Canada**  
Instructional Assistant

March 2023 – April 2023

- TUSK Upskilling Series: Machine Learning Software Foundations - Winter 2023, Spring 2023

## University of Toronto, Toronto, Canada

Graduate Teaching Assistant

- CSC413/2516: Neural Networks and Deep Learning - Fall 2025
- CSC457H1: Principles of Computer Networks - Fall 2020
- CSC343: Introduction to Databases - Spring 2021
- CSC263: Data Structures and Analysis - Fall 2021

## TECHNICAL SKILLS

---

**Languages:** Python (Expert) • Java • R • C • C++ • SQL (Familiar)

**NLP/AI Specializations:** LLMs • Transformers • Model Merging • Pruning • Distillation • Reward Modeling • Reinforcement Learning • Sequence Modeling • RNN • LSTM • GAN • Encoder-Decoder Models • LLM Safety and Security • Interpretability and Fairness • LLM Evaluation and Benchmarking • Language Modeling • Mixture of Agents • Mixture of Experts (MoE) • Retrieval-Augmented Generation (RAG) • Continual Pretraining • Multilingual Learning • Multimodal Learning • Automated Red-teaming • LLM Safeguards • LLM Pretraining and Post-training • Modular AI • Decentralized Training

**Packages:** PyTorch • Huggingface Transformers • Scikit-Learn • LangChain • LangGraph • Megatron-LM • vLLM • Meta Lingua • Alignment Handbook • Axolotl • NumPy • Pandas • SpaCy • NLTK • Matplotlib/Seaborn • BeautifulSoup • DataTrove • Open AI Agent SDK • Dify • Axolotl • LM Eval Harness • Stanza • Stanford NLP • Google Lens • Tesseract • LlamaParse • Unstructured

**Cloud/DevOps:** AWS (SageMaker, Trainium, Glue, EC2, ECR) • PySpark • ClearML • CircleCI • Runpod • Docker • Git • Weights and Bias • Claude Code • Langfuse • E2B Sandbox • MCP

## HONORS and AWARDS

---

- Faculty of Arts & Science Top (FAST) Doctoral Award, University of Toronto, 2025
- Vector PhD Research Grant, Vector Institute, 2025
- MITACS Accelerate Scholarship, University of Toronto, 2021
- Vector Scholarship in Artificial Intelligence, University of Waterloo, 2020 [DECLINED]
- Ranked 1<sup>st</sup> (Highest GPA) among all students of BSc in computer engineering entered in 2015, University of Tehran, Iran

## Volunteer Services

---

Community Support:

- Vice Chair, ACM-W Professional Chapter (since September 2025)
- Free Mentorship Program for Juniors in Tech (since 2023)
- MScAC Mentorship Program, Mentor (2023–2024 cohort)

Conference & Workshop Reviewing:

- Paper Reviewer, ICLR 2026 Workshop on Principled Design for Trustworthy AI - Interpretability, Robustness, and Safety across Modalities
- Paper Reviewer, NeurIPS 2025 (Position Paper Track)
- Paper Reviewer, ICLR 2025 Workshop on Modularity for Collaborative, Decentralized, and Continual Deep Learning
- Paper Reviewer, AAAI 2024 Workshop on Responsible Language Models

## Languages

---

- Farsi - Native proficiency
- English - Professional working proficiency
- Arabic - Limited working proficiency
- French - Limited working proficiency