Arash Asgari

Arashasg | asgari-arash | website | +1(437)210-3969
arash.asgari.m@gmail.com

Summary

Applied Machine Learning Engineer and Researcher with 3+ years of experience specializing in LLM evaluation, computer vision pipelines, and serverless MLOps. Backed by top-tier academic research (Vector Institute), with a proven track record of translating rigorous AI benchmarking into scalable, commercial applications across NLP and autonomous agents.

Work Experience

Machine Learning Researcher, York University May 2024 - August 2026

- **Designed and implemented a large-scale evaluation pipeline** to benchmark AI fairness and bias, devising novel statistical agreement metrics to cross-validate performance across 8 open-source LLMs (ranging from 2B to 9B parameters) and 7 distinct evaluation frameworks (this work is submitted to ACL 2026).
- **Engineered advanced inference-time and training-time debiasing algorithms** using representation engineering, successfully leveraging persona vectors and activation space steering to mitigate model bias.
- **Executed rigorous adversarial testing and safety benchmarking** for state-of-the-art LLMs in medical QA, utilizing baseline, jailbreak, and in-context debiasing prompts to evaluate model resistance and vulnerability (this work is published in AAAI symposium).

Research Intern, Vector Institute January 2025 - May 2025

- **Engineered a comprehensive evaluation pipeline** to benchmark state-of-the-art LLM knowledge editing methods (e.g., ROME, MEMIT, and MEND), working under the supervision of Dr. Frank Rudzicz.
- **Deployed and tested large language models** on high-performance compute clusters, optimizing scripts for memory-efficient inference and data processing.
- **Analyzed model performance and robustness** using medical QA datasets like MedSynth, evaluating the specificity and generalization of post-edit model weights.

Software Engineering Researcher, Sharif University of Technology October 2021 - April 2024

- **Curated and open-sourced DistilKaggle**, a massive-scale benchmark dataset comprising over 542,000 Jupyter notebooks, engineering an extraction pipeline to calculate 34 distinct structural and script-based code metrics (this work is published in MSR conference).
- **Engineered an automated code evaluation framework** by fine-tuning a DistilBERT transformer for semantic opinion mining, devising a novel metric that classified the code quality of 132,000+ computational notebooks with 89% accuracy (this work is published in Empirical Software Engineering Jouranl).
- Conducted rigorous **performance benchmarking of ML inference pipelines** using NVIDIA Multi-Instance GPU (MIG) architecture, analyzing hardware utilization, throughput, and latency across varying load profiles.

Machine Learning Engineer, Drivee, California, US February 2023 - January 2024

- **Architected a serverless image processing pipeline** using Google Cloud Functions in a pipe-and-filter architecture, sequentially automating exposure regularization, vehicle smoothing, and high-fidelity image enhancements.
- **Engineered a custom YOLO-based background removal model** and data workflows, powering the core AI engine for the Drivee iOS application to automatically generate studio-quality 360° virtual tours with custom dealership branding.
- **Developed an autonomous LLM-based conversational agent** using LangChain and OpenAI Agent frameworks, integrating directly with live dealership databases to deliver highly personalized, real-time vehicle recommendations and drive customer engagement.

Education

2024 - Aug 2026	M.Sc (Computer Science) at **York University**	(GPA: **4.0/4.0**)
2021 - 2024	M.Sc (Computer Engineering) at **Sharif University of Technology**	(GPA: 18.28/20.0)
2017 - 2021	B.Sc (Computer Engineering) at **Sharif University of Technology**	(GPA: 19.57/20.0)

Skills

LLM & Applied AI Frameworks	PyTorch, TensorFlow, Scikit-learn, HuggingFace, LLM Fine-tuning, RAGs, Agentic AI, MCP, Langchain, ADK (Agent Development Kit), Transformers
MLOps & Cloud	Docker, kubernetes, AWS (S3, EC2, and Redshift), CI/CD, model monitoring, Wandb, FastAPI, Linux, Bash Scripting
Data Platforms	Spark (PySpark, SparkML), Hadoop (MapReduce), Hive, SLURM
Languages	Python, R, JavaScript, SQL
Data Analysis & Vis	Pandas, NumPy, Matplotlib, Seaborn
Databases	MySQL, PostgreSQL, MongoDB

Publications

Asgari, Arash, Huan Wu, et al. (2026). "Quantifying Metric and Model Agreement in Bias Evaluation of Large Language Models". In: *The 64th Annual Meeting of the Association for Computational Linguistics*.

Asgari, Arash et al. (Nov. 2025). "MedPerturbing LLMs: A Comparative Study of Toxicity, Prompt Tuning, and Jailbreaks in Medical QA". In: *Proceedings of the AAAI Symposium Series* 7.1, pp. 438–447. DOI: 10.1609/aaaiss.v7i1.36916. URL: https://ojs.aaai.org/index.php/AAAI-SS/article/view/36916.

Ghahfarokhi, Mojtaba Mostafavi et al. (2025). "Predicting the understandability of computational notebooks through code metrics analysis". In: *Empirical Software Engineering* 30.4, p. 98.

Naziri, Amirreza, Arash Asgari, Aijun An, et al. (2025). "From Bias to Breakdown: Benchmarking Failure Mode Analysis of Single-cell RNA Sequencing Foundation Models in Acute Myeloid Leukemia". In: *Proceedings of the AAAI Symposium Series*. Vol. 7. 1, pp. 553–557.

Naziri, Amirreza, Arash Asgari, Eleftherios Sachlos, et al. (2025). "Improving Classification of Cell Types in Acute Myeloid Leukemia with Self-guided Masking Technique". In: *NeurIPS 2025 Workshop on AI Virtual Cells and Instruments: A New Era in Drug Discovery and Development*. NeurIPS, AI4D3.

Mostafavi Ghahfarokhi, Mojtaba et al. (2024). "DistilKaggle: a distilled dataset of Kaggle Jupyter notebooks". In: *Proceedings of the 21st International Conference on Mining Software Repositories*, pp. 647–651.

Projects

ParsSum: Transformer-Based Persian Summarizers Source Code on GitHub

Fine-tuned four near-SOTA Persian summarization transformers, curated two novel datasets, and deployed the end-to-end pipeline via a Django web app.

Persian News Sentiment Analysis Pipeline Source Code on GitHub

Developed a custom NLP sentiment classification pipeline for news stories. Engineered and evaluated two distinct architectures: a fine-tuned Persian BERT and a custom-built RNN.

Additional Projects View More on GitHub

Explore additional natural language processing and computer vision projects, including LLM fairness evaluation pipelines, dataset curation (DistilKaggle), and medical image segmentation.

References

Dr. Saeid Ghafouri
Machine Learning Systems Researcher, Huawei
saeid.ghafouri1@huawei.com

Dr. Laleh Seyyed Kalantari
Professor & Supervisor, York University
lsk@yorku.ca