Faridoun Mehri

feraidoonmehri@gmail.com | github.com/NightMachinery | linkedin.com/in/feraidoon-mehri | Scholar

Research Interests

My current work focuses on **Transformer interpretability**, where I've developed SOTA white-box **attribution** techniques (validated on *ViTs*, extensible across modalities) and formalized them in the *LibraGrad* framework. Moving forward, I aim to leverage attribution methods to enhance model capabilities, particularly in improving robustness to *distribution shifts*, *spurious correlations*, and *adversarial attacks*—having specialized coursework and a strong interest in the latter. I also plan to extend *LibraGrad* to **SSMs** (*Mamba*). Beyond attribution, I am especially interested in **mechanistic interpretability**, utilizing techniques like *sparse autoencoders* and *activation patching* to understand the internals of models. Impressed by my daily usage of **LLMs**, I'm eager to contribute to areas such as **reasoning**, **AI alignment**, **evaluation**, **prompt engineering**, **modular deep learning**, and **agents**.

Publications

1. **Faridoun Mehri**, M. S. Baghshah, and M. T. Pilehvar, "LibraGrad: Balancing Gradient Flow for Universally Better Vision Transformer Attributions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025 (**Oral** Presentation, Acceptance Rate: **0.74%**, Review Scores: 5, 5, 4 out of 5) [paper]
2. **Faridoun Mehri**, M. Fayyaz, M. S. Baghshah, and M. T. Pilehvar, "SkipPLUS: Skip the First Few Layers to Better Explain Vision Transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (Oral TCV@CVPR)*, pp. 204–215, June 2024 [paper | slides]

Education

Sharif University of Technology — Tehran, Iran
Master of Science (M.Sc.) in AI & Robotics — *9/2022 - ongoing*

- Supervisors: *Dr. Mahdieh Soleymani Baghshah* and *Dr. Mohammad Taher Pilehvar*
- Thesis: Interpretation of Transformer Models
- GPA: **19.7**/20.0
- Rank: **1st**/35
- **Courses**: (The grades are from 20.)
  Deep Learning (20) | *Convex Optimization* (20) | *Machine Learning* (20) | *Natural Language Processing* (20)
  DSP (19.7) | *Security & Privacy in Machine Learning* (19.3) | *Information Theory* (18.5) | *Signals & Systems* (20)

Sharif University of Technology — Tehran, Iran
Bachelor of Science (B.Sc.) in Computer Science — *9/2016 - 7/2021*

- GPA: **17.1**/20.0 (Average Department GPA for 2016 CS Cohort: 16.34)
- GPA of the Last 64 Credits: **19.0**/20.0

Shahid Beheshti (National Organization for Development of Exceptional Talents) — Sabzevar, Iran
Pre-University & High School Diploma in Mathematics & Physics — *9/2012 - 7/2016*

- Pre-University GPA: **20.0**/20.0
- High School GPA: **19.9**/20.0

Language Proficiency

• **English** TOEFL iBT: **117**/120 (R:**30**, L:**30**, S:**28**, W:**29**) • **Persian**: Native • **Arabic**: Basic

Honors and Awards

- Awarded the ML Safety Student Scholarship of 2023 from the *Center for AI Safety*
- Ranked **second** (out of 16,703 students) in the Iranian national graduate entrance exam in *AI & Robotics* (and all other CE majors) of 2022 (**99.99$^{\text{th}}$ percentile**)
- Ranked **fourth** (out of 1,322 students) in the Iranian national graduate entrance exam in *Computer Science* of 2022 (**99.70$^{\text{th}}$ percentile**)
- Ranked **sixth** (out of 115,803 students) in the Iranian national undergraduate entrance exam in *Foreign Language Studies* of 2016 (**99.99$^{\text{th}}$ percentile**)
- Ranked 331st (out of 162,731 students) in the Iranian national undergraduate entrance exam in *Mathematics & Physics* of 2016 (**99.80$^{\text{th}}$ percentile**)

Research Experience

Interpretation of Transformer Models 09/2022 - ongoing
Machine Learning Lab (MLL) *Sharif University of Technology*
Under the supervision of *Dr. Mahdieh Soleymani Baghshah* and *Dr. Mohammad Taher Pilehvar*

- See Research Interests, Publications, or the Extended CV

Blockchain-Based Solutions to Privacy-Preserving Health Data 08/2022 - 09/2022
Under the supervision of *Dr. Parviz Rashidi Khazaee*'s student, *Amin Samsami* *Urmia University of Technology*

- **health_blockchain** (Python): prototyped a blockchain for storing health data privately in a distributed manner

Manifold Learning and High-Dimensional Clustering 12/2021 - 09/2022
Sharif Optimization and Applications Laboratory (SOAL) *Sharif University of Technology*
Under the supervision of *Dr. Amir Daneshgar*, *Dr. Mohammad-Hadi Foroughmand*, and *Dr. Mojtaba Tefagh*

- Solved the Optimizer 2022 challenges around manifold learning and clustering in high-dimensional data with outliers and noise
- Visualized the high-dimensional input data, the detected manifolds, clusters, convex hulls, outliers, and noise, which was critical in diagnosing many bugs
- Developed, tested, and calibrated the automatic judge (the autograder) of the Optimizer 2022 challenges
- Designed and tested the data generation algorithms for the Optimizer 2022 challenges
- Created a modular benchmark system for clustering algorithms that measures memory usage, execution time, and various accuracy metrics, with support for big (~100GB) data (Dask, RAPIDS, scikit-learn)

Teaching Experience

Workshop Instructor
LLM Workshops, *Clustering in Python from Scratch* *WSS & Sharif University (2022-2024)*

Head Teaching Assistant
Computer Networks, Principles of Computer Systems *Sharif University of Technology (2022-2023)*
Teaching Assistant Graduate level courses are marked with $^+$.
Large Language Models$^+$, Deep Learning$^+$, NLP$^+$, AI *Sharif University of Technology (2018-2023)*
Big Data Engineering, Advanced Programming, Digital Logic Design

Skills

Programming Languages I Use Frequently: Python, Zsh (shell scripting), Elisp
Programming Languages I Have Written Some Useful Things in: Julia, Java, Common Lisp, Golang, Perl, Clojure, Scala, Kotlin, Racket, Lua, Javascript, Node.js, SQL, VB.NET, C#, C++, Rust
ML/Data Libraries: PyTorch, HuggingFace, timm, Google's JAX, Flux.jl, scikit-learn, Nvidia's Rapids, conda, numpy, pandas, einops, Matplotlib, seaborn, plotly
Backend Technologies: GNU/Linux, Docker, Caddy, Akka, Redis, FastAPI
Developer Tools: Git, tmux, Emacs, vim, VSCode, Jupyter
Other Technical Skills: LaTeX, profiling, web scraping, blockchains, distributed systems, regex, documentation writing and note-taking, prompt engineering

Notable Open-Source Projects

- Reported hundreds of bugs (You need to sign in to Github before viewing this link.)
- Contributed to **HuggingFace Pytorch Image Models** (`timm`), **HuggingFace Datasets**, **sioyek**, **ugrep**, **Emacs**, **Doom Emacs**, **Flux.jl**, **Zsh**, and other FOSS projects
- **stochastic** (Julia): an infectious disease model (a grad course project of mine), a Poisson picture redrawing filter, a colorful animator of a 2D ising model, and more
- **twitter-scraper** (Python, Zsh, Docker): a fault-tolerant, distributed Twitter scraper which stores the data in Neo4j
- **distributed-prime-generator** (Scala, Akka, Docker): a fault-tolerant, distributed prime number generator
- **price_detector_fa** (Python, Hazm): extracts product/price/amount tuples from Persian text using rule-based methods
- **brish**: Python-Zsh integration library enabling safe variable interpolation and parallel execution (**83k+** downloads)

Extended Version

- The unabridged, up-to-date version of this CV is available at files.lilf.ir/CV.pdf