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
- **Notable Courses**: (Graduate level courses are marked with $^+$. The grades are from 20.)
  Advanced Programming (20) | *Artificial Intelligence* (19.7) | *Probability* (19.8) | *Statistics* (20) | *Applications of Stochastic Processes*$^+$ (19) | *Data Transfer & Networks* (19.3) | *Theory of Computation & Complexity* (18.4) | *Design of Programming Languages* (18) | *Mathematical Logic* (18.5) | *Systems Theory* (18.6) | *Principles of Computer Systems* (18.9) | *Cryptography, Distributed Systems, & Blockchains*$^+$ (18.9) | *Big Data Engineering* (19) | *Analysis of Algorithms* (20) | *Engineering Mathematics* (20)

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 (**99.99th percentile**)
- Ranked **fourth** (out of 1,322 students) in the Iranian national graduate entrance exam in *Computer Science* of 2022 (**99.70th percentile**)
- Ranked **sixth** (out of 115,803 students) in the Iranian national undergraduate entrance exam in *Foreign Language Studies* of 2016 (**99.99th percentile**)
- Ranked 331st (out of 162,731 students) in the Iranian national undergraduate entrance exam in *Mathematics & Physics* of 2016 (**99.80th 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*

- Synthesizing different gradient-based attribution methods into a single theoretical framework
- Proposed *LibraGrad*, a theoretically grounded post-hoc approach that corrects gradient imbalances through pruning and scaling of backward paths, without changing the forward pass or adding computational overhead
  - ∗ Universally enhanced all gradient-based attribution methods while outperforming existing white-box methods across 8 architectures, 4 model sizes, and 4 datasets on faithfulness, completeness error, and segmentation AP
  - ∗ Showcased unmatched qualitative capabilities through precise text-prompted region highlighting in CLIP models and accurate class discrimination between co-occurring animals in ImageNet models
- Pioneered *PLUS*/*SkipPLUS*, state-of-the-art Transformer attribution methods designed for universal composability with existing methods
  - ∗ Proposed *FullGrad+* and *XGradCAM+*, with *LibraGrad* elevating *FullGrad+* to SOTA
  - ∗ Isolated, generalized, and simplified recent advances in Transformer attribution into *PLUS*
  - ∗ Quantified that *PLUS* surpasses *Rollout*: achieving higher faithfulness and segmentation AP while offering better composability, greater robustness, lower complexity, and faster inference speed
  - ∗ Engineered comprehensive evaluation pipeline for Transformer attribution methods, implementing methods from scratch and developing novel visualization techniques
- Designed *SeCA*, a method to enhance the class-discriminativity of attribution methods and enable class aggregation (many fine-grained classes into a single coarse-grained class)

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.)

Popular FOSS Projects I Have Contributed To

- **HuggingFace Pytorch Image Models** (AKA `timm`, Python): fixed bugs
- **HuggingFace Datasets** (Python): added features
- **sioyek** (C++): fixed bugs and added features
- **ugrep**: suggested improvements which *Dr. Robert van Engelen* liked and subsequently implemented
- **Emacs** (Elisp): added features
- **Doom Emacs** (Elisp): added features and fixed bugs
- **Flux.jl** (Julia): fixed mistakes in the documentation and wrote more documentation
- **Zsh**: reported a bug which was promptly fixed
- **fzf-tab** (Zsh): fixed bugs
- **learnxinyminutes.com**: fixed mistakes
- **bkmeans** (Python, scikit-learn): fixed bugs
- **Anime4KCPP** (C++): added macOS support

Academic 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 (a distributed graph database), plus a high-level CLI API for querying the data and load testing the system
- **distributed-prime-generator** (Scala, Akka, Docker): a fault-tolerant, distributed prime number generator using the actor model
- **MLP From Scratch** (Python, numpy): a simple trainable MLP using only numpy with support for batch axes
- **random-shuffle-SGD** (Python, PyTorch): an implementation of the paper "How Good is SGD with Random Shuffling?"
- **price_detector_fa** (Python, Hazm): extracts product/price/amount tuples from Persian text using rule-based methods
- **Cross-Lingual Transfer Learning From English to Persian With Zero-Shot/Few-Shot MAD-X** (Python, PyTorch, HuggingFace, adapter-transformers)
- **Char-RNN** (Python, PyTorch): a character-level language model using GRUs and beam search decoding
- **fanfiction-classifier** (Python, JAX, Haiku, Optax): a character-level, variable-length text classifier using (optionally dilated) convolutional layers, dropout, layer normalization, and learning rate scheduling (runs on both TPUs and GPUs)
  - A similar model in Julia's Flux.jl
- **reo** (Racket): a toy DSL
- **Toy MIPS CPU**: an implementation of a simple, non-pipelining CPU using gate-level Verilog

My Own FOSS Projects

- **brish**: a thread-safe Python library using the metaprogramming API which lets the user embed and run Zsh code in Python via parallel processes, supporting safely interpolating Python variables into the Zsh code (**83k+** downloads)
- **readability-cli** (Node.js): declutters and sanitizes scraped HTML using Mozilla Readability
- **betterborg** (Python): a pluggable Telegram (user)bot based on Telethon (forked from uniborg)
  - Integration with my library brish, to allow using Telegram as, essentially, a terminal emulator, with support for safely giving users limited access to specific Unix commands

    - A time and habit tracker that supports hierarchical activities, with an always learning DSL to submit events and request visualizations
    - A declarative DSL based on JSON for Telegram's inline query UI
    - A prototype of a declarative DSL based on JSON for Telegram's bot UI
- **JupyterGarden** (Python, FastAPI): an HTTP REST API to run code in Jupyter kernels, for languages with expensive startup costs
- **possiblycat** (Golang): cat with a timeout on waiting for the first byte of stdin
- **prefixer** (Golang): a modern alternative to GNU cut
- **jalalicli** (Golang): a CLI utility for Jalali (Shamsi) dates
- **jalali-calendar-cli** (Python, Perl): a TUI Jalali (Shamsi) calendar (holiday data extracted using LLMs from official PDFs)
- **My Megarepo of Scripts**
    - Prompt crafting tools
    - Scraping tools (or API wrappers) for github.com, semanticscholar.org, arxiv.org, edu.sharif.edu, cw.sharif.ir, goodreads.com, reddit.com, tumblr.com, spotify.com, store.steampowered.com, nationalgeographic.com, bing.com, duckduckgo.com, fanfiction.net, kitsu.io, myanimelist.net, techcrunch.com, sounds-resource.com, lesswrong.com, messaged.com/tldr, web.archive.org, techmeme.com, news.ycombinator.com, sanjesh.org, patreon.com, . . .
        - **web2audio**: creates an audiobook from a given set of URLs using pretrained deep TTS models
        - **t.me/techmemenews**
        - **t.me/tldrnewsletter**
        - **t.me/sharif_edu_diff**: notifications of changes in the scheduling, capacity, instructors, etc. of SUT's courses of the current semester
        - **sharif_course_list**: courses offered by SUT, in HTML and JSON, under git for archival purposes
        - **ddg2json** (Python): parses scraped HTML of DuckDuckGo pages into JSON
        - **r_rational**: the subreddit r/rational archived in plain-text org-mode (good for, e.g., doing offline full-text searches)
    - A keyboard controller for mouse clicks and drags which allows one to use mouse-only software efficiently (Hammerspoon, Lua)
    - A general Zsh function memoizer using Redis
    - A git-backed reminder system supporting natural language for setting the due time, recurrent reminders, and integration with Telegram, Google Calendar, and macOS/iOS (notifications, widgets, wallpapers)
    - A Redis-backed RSS manager with integrations for Telegram and Amazon Kindle (supports podcasts, as well)
    - A recurrent bill manager that parses my org-mode (plain-text) notes and presents me the bills likely to be due
    - A whole suite of index-less information retrieval CLI tools for searching IRC logs, reminders, contacts, function definition locations, oft-used directories, music, ...
        - A Perl-based custom grep tool for tree-shaped documents (org-mode), plus a TUI for viewing the results (built on top of Emacs)
    - A file tagging system using the file names as the database
    - A "thermostat" for display brightness based on the brightness of the current screen content
    - A clipboard manager integrated with Zsh and Emacs
    - Integration with Supercollider, to allow using real-time, stochastic generative audio for auditory notifications in CLI applications
- **vcard-to-json** (Clojure): a CLI tool to convert vCard files into JSON
- **rtl_reshaper_rs** (Rust): a CLI tool to reshape and reorder bi-directional, Arabic/Persian text for display in apps that do not support them natively

FOSS Projects I Developed When I Was a Child 2007-2012

- **Aero Form** (VB.NET): A subclass of Form, it allowed the user to extend the Aero effect of Windows Vista to the whole window (or a subset thereof). It was downloaded **60k** times on marketplace.visualstudio.com.
- **HyperAero Form** (VB.NET): The much-upgraded version of the above, it was my first lesson in marketing; while it was better than Aero Form for virtually all purposes, its less catchy name doomed it to obscurity. It was downloaded **13k** times on marketplace.visualstudio.com.
- **Notify Msg** (VB.NET): Probably my first useful library, it allowed one to show desktop notification popups, supporting images and other goodies. It was downloaded **1.5k** times.
- **File Splitter** (VB.NET): A WPF-powered GUI for splitting and merging files.
- **Animation Maker** (VB.NET): Inspired by Windows Presentation Foundation's ease of creating animations, I built an animation library for Windows Forms using the reflection API (demo).

References

- **Dr. Mahdieh Soleymani Baghshah** (H-Index: 23, Scholar)
  Associate Professor, Department of Computer Engineering, Sharif University of Technology
  MSc Thesis Supervisor
  Email: soleymani@sharif.edu
- **Dr. Mohammad Taher Pilehvar** (H-Index: 33, Scholar)
  Senior Lecturer, School of Computer Science, Cardiff University; *Affiliated Lecturer*, University of Cambridge
  MSc Thesis Supervisor
  Email: pilehvarmt@cardiff.ac.uk
- **Dr. Mohammad Hadi Foroughmand Aarabi** (H-Index: 5, Scholar)
  Assistant Professor, Department of Mathematical Sciences, Sharif University of Technology
  BSc Research Supervisor
  Email: foroughmand@sharif.edu
- **Dr. Mojtaba Tefagh** (H-Index: 5, Scholar)
  Assistant Professor, Department of Mathematical Sciences, Sharif University of Technology
  BSc Research Supervisor
  Email: mtefagh@sharif.edu