MARYAM REZAEE

ms.maryamrezaee@gmail.com
github.com/msmrexe
maryamrezaee.me

I am a computer science researcher working across Explainable Artificial Intelligence (XAI), reasoning in large language models, & cognitively-inspired AI. I am driven by the vision of taming the neural black box by introducing structure into generative systems, aiming to make AI an interpretable, verifiable tool that better emulates human thought. After working on concept-level attribution to examine generation behaviour, I have begun exploring neurosymbolic architectures that leverage structural inductive biases to curb hallucinations and enforce consistency. Ultimately, I hope to improve the fluency and trustworthiness of neural networks.

Education

M.Sc. in Computer Science — Sep 2023 – Aug 2026

Department of Mathematical Sciences, Sharif University of Technology — AVG: 18.52 / 20.00

Ranked 8th nationwide in the National University Entrance Exam for Computer Science

Thesis: *Interpretability in Generative Models: Investigating the Mechanisms Behind Output Generation in Large Language Models*

B.Sc. in Computer Science — Sep 2019 – Sep 2023

Department of Mathematics and Statistics, University of Isfahan — AVG: 16.66 / 20.00

Project: *Understanding the Application of Machine Learning in Cell Signalling Pathway Analysis* — RANK: 13 / 56

Research Experience

Exploring Neurosymbolic AI and Cognitive Architectures — Aug 2025 – Present

Sharif University of Technology, Tehran, Iran

- Conducting a literature review on the integration of NNs with symbolic reasoning systems to inform future research.
- Studying key topics including program synthesis, logic-based model integration, Computational Narratology, and the role of human inductive biases (specifically Active Inference and the FEP) in cognitive architectures.

Arc-CBM: Prototype-Grounded Concept Bottlenecks with Manifold-Grounded Visual Auditing for Medical Image Analysis — May 2025 – Present

Mostafa Abdolmaleki, Maryam Rezaee, Fatemeh Seyyedsalehi

Trustworthy and Generative Machine Learning (TGML) Lab, Sharif University of Technology, Tehran, Iran

- Developed a prototype-grounded concept bottleneck model that defines clinical concepts using L2-projected positive and negative prototypes derived from report-labeled support images.
- Designed an attribution method that projects prototype-margin gradients onto token-level tangent spaces to provide localized visual explanations without extra encoder passes.

Concept-Based Interpretability for Retrieval-Augmented Generation (RAG) Systems on Large Language Models — May 2025 – Present

TGML Lab, Sharif University of Technology, Tehran, Iran

- Mentoring a B.Sc. student in developing a novel framework for RAG interpretability.
- Developing methods using Concept Activation Vectors and Automatic Concept-based Explanations to enhance the transparency and traceability of RAG outputs.

Sentence-Level Attribution for Black-Box Large Language Model Outputs via Energy-Based Surrogates

Feb 2025 – Jan 2026 Link

Maryam Rezaee, Pooriya Safaei, Maryam Asgarinezhad, Fatemeh Seyyedsalehi
TGML Lab, Sharif University of Technology, Tehran, Iran

- Proposed a model-agnostic, post-hoc framework (ESCI) to interpret black-box LLMs accessed only via APIs.
- Designed an energy model to quantify the conceptual links between user prompts and LLM-generated responses.
- Trained an efficient interpreter network to provide sentence-level importance scores.
- Demonstrated that this efficient interpreter can compete with the attribution fidelity of baselines.
- *Status: preprint available on arXiv; final version being submitted to TNNLS.*

Interpretability in Generative Models: Investigating the Mechanisms Behind Output Generation in Large Language Models

Sep 2024 – Aug 2026 Link

Master's Thesis Researcher under the supervision of Dr. Fatemeh Seyyedsalehi
TGML Lab, Sharif University of Technology, Tehran, Iran

- Conducted a comprehensive review of intrinsic, post-hoc, concept-based, & mechanistic interpretability methods and EBMs in NLP and categorised available methods in a detailed literature review.
- Developed a post-hoc framework to interpret instance-based input-output attribution in LLM generation.

Understanding the Application of Machine Learning in Cell Signalling Pathway Analysis and Developing a DEG Analysis Handbook

Feb 2023 – Sep 2023 Link

Bachelor's Project Researcher under the supervision of Dr. Fatemeh Mansoori
University of Isfahan, Isfahan, Iran

- Conducted a literature review on the application of ML and DL (e.g., SVMs) in bioinformatics.
- Analysed and documented methodologies for Gene Expression and Differentially Expressed Gene (DEG) analysis.
- Authored a technical handbook and implementation roadmap for future students to use in DEG analysis.

A Neurosymbolic Architecture for Coherent Narrative Generation

Jan 2021 | 2025 Link

Undergraduate Researcher under the supervision of Dr. Jaafar Almasizadeh
University of Isfahan, Isfahan, Iran

- Designed a Neurosymbolic "Director-Actor" architecture to improve LLM coherence, operationalising narrative structures via Signal Processing (Gaussian envelopes), Gated FSMs, and A* Search.
- Modelled character archetypes as Vector Centroids to enforce consistency and framed narratives as cognitive data compression schemas optimising for Active Inference.

Foundational Research in Cognitive Neuroscience

Aug 2016 – Feb 2017

Team Researcher (1st Kashan Neuroscience Workshop for Students)
Kashan University of Medical Sciences, Kashan, Iran

- Achieved 2nd Place in a multi-stage research competition for a portfolio on neuroscience and psychology.
- Researched and presented on cognitive topics, including the neural basis of creativity (divergent thinking), hypnotic analgesia, and psychological models of learning/cognitive styles.
- Conducted foundational reviews on core neuroanatomy (Limbic System), the brain-mind problem, and applied clinical neuroscience as part of the workshop curriculum.

Three Centuries in Pursuit of the Theory of Everything

Oct 2015 – Feb 2016

Independent Researcher (Khawrazmi Youth Award)
Farzanegan School (NODET: National Organization for Development of Exceptional Talents), Kashan, Iran

- Awarded 2nd Place in the Physics category at the Khawrazmi Youth Award (Kashan regional).
- Authored a comprehensive research paper on the historical and theoretical development of a unified field theory in physics, analysing key milestones from classical mechanics to modern physics.

Teaching Experience

Relevant to Academic Major

- **Head Teaching Assistant: *Generative Models*** *(M.Sc.)* — Link — Sep 2025 – Feb 2026
  Dr. Fatemeh Seyyedsalehi, Sharif University of Technology
- **Teaching Assistant: *Machine Learning Operations*** *(B.Sc.)* — Sep 2025 – Feb 2026
  Dr. Fatemeh Seyyedsalehi, Sharif University of Technology
- **Head Teaching Assistant: *Machine Learning*** *(B.Sc.)* — Sep 2025 – Feb 2026
  Dr. Mohsen Alambardar, University of Isfahan
- **Instructor: *Fundamentals of Technological Product Development*** — Jul 2024 – Sep 2024
  Tehran University of Medical Sciences
- **Head Teaching Assistant: *Advanced Programming*** *(B.Sc. Math & CS)* — Jan 2023 – Jun 2023
  Dr. Fatemeh Mansoori, University of Isfahan
- **Head Teaching Assistant: *Operating Systems*** *(B.Sc.)* — Jan 2023 – Jun 2023
  Dr. Mojtaba Rafiee, University of Isfahan
- **Instructor: *Algorithms and Data Structures*** *(B.Sc.)* — Nov 2022 – Feb 2023
  Private Classes, University of Isfahan
- **Instructor: *Python Programming*** *(Elementary to Advanced)* — Sep 2019 – Feb 2023
  Private Classes, University of Isfahan

Additional Topics

- **Instructor and Mentor: *Graphic Design with Adobe Illustrator*** — Jan 2023 – Jun 2023
  Academic Association of Mathematics and Computer Science (AMCSUI), University of Isfahan
- **Instructor: *English Language*** *(Intermediate to Advanced)* — Sep 2016 – Nov 2023
  Private Classes

Selected Course Projects

- **System 2 AI** (M.Sc.) — Feb 2025 – Jun 2025
  Dr. M.H. Rohban & Dr. M. Soleymani, Sharif University of Technology — Audit
  PY Program Generation w/ Seq2Seq Models · PY Symbolic Regression · PY GraphRAG · PY LLM Agent
- **Deep Learning** (M.Sc.) — Feb 2025 – Jun 2025
  Dr. Fatemeh Seyyedsalehi, Sharif University of Technology — Mark: 20.00 / 20.00
  PY GPT2 · PY Math Reasoning in LLMs · PY PINN · PY Atari DQN · PY SSL & Contrast. Learning · PY FNN from Scratch

Generative Models (M.Sc.) Sep 2024 – Feb 2025
Dr. Fatemeh Seyyedsalehi, Sharif University of Technology Mark: 18.00 / 20.00
PY RNN vs Transformer · PY Diffusion/DDPM · PY EBM · PY β-VAE · PY Pix2Pix · PY Norm. Flows · PY Bayesian Network

Machine Learning (M.Sc.) Sep 2024 – Feb 2025
Dr. Ali Sharifi Zarchi, Sharif University of Technology Mark: 20.00 / 20.00
PY Neural Decompo. · PY LoRA from Scratch · PY Word2Vec · PY Knowl.Distil. (KD) & CLIP · PY MobileNet, Eff. & KD.

Matrix Computations (M.Sc.) Sep 2023 – Feb 2024
Dr. Mohammad Reza Razvan, Sharif University of Technology Mark: 19.00 / 20.00
PY SVD RGB-Image Compressor from Scratch · MATLAB Comparing Linear Solvers Implemented from Scratch

Computer Networking (B.Sc.) Feb 2023 – Jun 2023
Dr. Mohsen Alambardar, University of Isfahan Mark: 18.75 / 20.00
PY Group Chat Application with TCP Protocol and Highspeed Data Sharing

Artificial Intelligence (B.Sc.) Sep 2022 – Feb 2023
Dr. Fatemeh Mansoori, University of Isfahan Mark: 19.50 / 20.00
PY AI Search Algorithms for Solving Mazes · PY MDP Gridworld Solver with Value & Policy Iteration

Cryptography (B.Sc.) Sep 2022 – Feb 2023
Dr. Mojtaba Rafiee, University of Isfahan Mark: 19.50 / 20.00
Feasibility of Proving Authenticity and Data Integrity for P2P Using Hash

Fundamentals of Operating Systems (B.Sc.) Feb 2022 – Jun 2022
Dr. Mojtaba Rafiee, University of Isfahan Mark: 19.50 / 20.00
PY Reader-Writer Lock Implementation with Three Priorities

Mathematical Databases (B.Sc.) Feb 2022 – Jun 2022
Dr. Fatemeh Mansoori, University of Isfahan Mark: 19.60 / 20.00
PostgreSQL University Course Registration System · PostgreSQL Academic Transcript Generator & Analyzer

Mathematical Software (B.Sc.) Feb 2022 – Jun 2022
Dr. Najmeh Hoseini, University of Isfahan Mark: 17.75 / 20.00
MATLAB Audio-Signal Analysis · MATLAB CLI University Portal for Grade Management

Algorithms & Data Structures (B.Sc.) Feb 2022 – Jun 2022
Dr. Jaafar Almasizadeh, University of Isfahan Mark: 20.00 / 20.00
PY Chinese Postman · PY Huffman Compressor · PY Max-Flow Altered · PY Prime Sieves Anal. · PY Topol. Sort Anal.

Advanced Programming (B.Sc.) Feb 2020 – Jun 2020
Dr. Jaafar Almasizadeh, University of Isfahan Mark: 20.00 / 20.00
PY RSA Cryptosystem Implementation from Scratch · PY CLI Banking System

Selected Talks

ReAGent: A Model-Agnostic Feature Attribution Method for LMs — Link — Aug 2025
Trustworthy & Generative ML (TGML) Lab, Sharif University of Technology

Interpreting Language Models with Contrastive Explanations — Link — Jun 2025
TGML Lab, Sharif University of Technology

Propagating Universal Perturbations to Attack LLM Guard-Rails — Link — May 2025
Deep Learning Seminars, Sharif University of Technology

Universal and Transferable Adversarial Attacks on Aligned LMs — Link — May 2025
Deep Learning Seminars, Sharif University of Technology

An Exploration of Concept-Based Interpretability Methods — Link — Mar 2025
TGML Lab, Sharif University of Technology

DiT: Scalable Diffusion Models with Transformers — Link — Feb 2025
Generative Models Seminars, Sharif University of Technology

Saliency-Guided Training: Interpretability and Robustness in DNNs — Link — Jan 2025
Machine Learning Seminars, Sharif University of Technology

Model Sparsity Can Simplify Machine Unlearning — Link — Dec 2024
Machine Learning Seminars, Sharif University of Technology

A Foray into Bioinformatics: ML in Cell Signalling Pathway Analysis — Link — Jun 2023
Bachelor's Project Presentation, University of Isfahan

Key Takeaways from the 19th Iranian Conference on HPE — May 2018
Invited Speaker, Post-Conference Seminar, Kashan University of Medical Sciences

Technical Skills

Areas of Expertise

- ML/DL
- GMs/LLMs
- XAI
- NeSy AI
- Reasoning
- RAG
- MLOps
- Bioinform
- CogNeuro

Languages / Core Tools

- Python
- R
- MATLAB
- PostgreSQL
- LaTeX
- JavaScript
- HTML/CSS
- Sockets
- Git/Bash

Modules / Frameworks

- PyTorch
- TensorFlow
- Scikit-learn
- NumPy
- Pandas
- Transformers
- Diffusers
- Threading
- RISC-V

Technical Software

- Illustrator
- InDesign
- Photoshop
- Premiere Pro
- After Effects
- Audition
- Figma
- Canvas
- Docker

Leadership & Project Management

Conceptual Prototype Designer Jul 2024 – Aug 2024
"Siba," the Interprofessional Gamified Platform for Intelligent Education and Evaluation of Medical, Nursing, and Pharmacy Students | *16th Scientific Olympiad for Medical Science Students of Iran*

Founder, Manager, Lead Designer May 2023 – Mar 2024
"PsyCity," the National Competitive and Educational Game for Students of Computer Science

Committee Representative Apr 2023 – May 2024
Assembly for the Academic Associations | *University of Isfahan*

Association President, Head of Content Creation Branch Oct 2022 – May 2024
Association for Mathematics & Computer Science (AMCSUI) | *University of Isfahan*

Content Lead Mar 2022 – Apr 2022
"Varpharin," the Gamified Event for Diabetic Patient Management with an Interprofessional Approach | *AMEE Grant Recipient, Presented at AMEE 2023 (Short Communication)*

Professional Activities & Service

Problem Setter and Instructor Dec 2025
National Artificial Intelligence Olympiad Committee of Iran | *Natural Lang. Process. & Comp. Vision*

Invited AI Panelist Aug 2025
International AMEE 2025 Conference, Live | *Symposia on AI in Health Professions Education*

English Interpreter & Student Task Force May 2023
24th Iranian Conference on Health Professions Education | *International Guests & Presenters*

UI/UX & Graphic Designer Since 2018
Freelance | *Academic & Institute Projects*

Academic & Technical Translator Since 2016
Freelance | *Book & Article Publications*

Awards

Honouree: **Tech & Innovation** 2024
17th International Harekat Festival, Iran

Top: **Digital Content** 2024
17th International Harekat Festival, Iran

Top: **Best Science Assoc.** 2024
17th University-Level Harekat Festival, UI, Iran

Top: **Best & Most Creative Pres.** 2024
17th University-Level Harekat Festival, UI, Iran

1st: **Most Active Association** 2024
Tadaee Festival, University of Isfahan, Iran

Honouree: **Tech & Innovation** 2024
Roshana Festival, Science Ministry, Iran

Honouree: **Educational Assoc.** 2023
16th International Harekat Festival, Iran

Honouree: **Best Team** 2018
Avicenna Student Creativity Cup, Iran

2nd: **Best Team** 2016
1st Kashan Neuroscience Workshop, Iran

2nd: **Best Research in Physics** 2016
18th Khawrazmi Youth Award, Kashan, Iran

1st: **Poetry** 2016, 2017, 2018
Regional Cultural-Artistic Cup, Kashan, Iran

1st: **Short Story** 2016, 2017
Regional Cultural-Artistic Cup, Kashan, Iran

Languages

- Full Proficiency in ***English***
- Native Fluency in ***Persian (Farsi)***