MARYAM REZAEE

✉ ms.maryamrezaee@gmail.com
🌐 msmaryamrezaee.github.io
github.com/msmrexe

A computer science researcher specializing in LLM Interpretability and Cognitively-Inspired AI. My objective is to develop robust, explainable systems by investigating the integration of human inductive biases and symbolic reasoning structures into neural models. My foundational experience includes a thesis on LLM input-output interactions and research in concept-based interpretability. I am now expanding this focus toward neurosymbolic architectures, evidenced by projects in program generation. I seek to advance this research with the aim of developing models capable of more robust and human-aligned reasoning.

Education

M.S. in Computer Science Sep 2023 – Present
Sharif University of Technology, Tehran, Iran GPA: 18.57 / 20.00
Thesis: *Interpretability in Generative Models: Investigating the Mechanisms Behind Output Generation in Large Language Models*

B.S. in Computer Science Sep 2019 – Sep 2023
University of Isfahan, Isfahan, Iran GPA: 16.66 / 20.00
Project: *Understanding the Application of Machine Learning in Cell Signalling Pathway Analysis*

Research Experience

Exploring Neurosymbolic AI and Cognitive Architectures Jul 2025 – Present

Independent Researcher
Sharif University of Technology, Tehran, Iran

- Conducting a self-directed literature review on the integration of neural networks with symbolic reasoning systems to inform future research.
- Studying key topics including program synthesis, logic-based model integration, and the role of human inductive biases in cognitive architectures.
- Implemented a seq2seq model for program generation as a practical exploration of neurosymbolic principles.

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

Research Mentor under the supervision of Dr. Fatemeh Seyyedsalehi
Trustworthy and Generative Machine Learning (TGML) Lab, Sharif University of Technology, Tehran, Iran

- Mentoring a B.S. student in developing a novel framework for RAG interpretability.
- Investigating the application of Concept Bottleneck Models (CBMs) to explain the internal generation processes.
- Developing methods using Concept Activation Vectors (CAV) and Automatic Concept-based Explanations (ACE) to enhance the transparency and traceability of RAG outputs.

A Concept Level Energy-Based Framework for Interpreting Black-Box Large Language Model Responses

Link Feb 2025 – Sep 2025

Graduate Researcher under the supervision of Dr. 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 a standalone, efficient interpreter network to provide sentence-level importance scores, mitigating common LLM biases without requiring further API queries.
- **Status:** *Under review at ICLR 2026*.

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

Sep 2024 – Present

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, and mechanistic interpretability methods for generative models, with a primary focus on LLMs.
- Developed a post-hoc framework to interpret instance-based input-output attribution in LLM generation.
- Analyzing internal model mechanisms and their influence on output generation.
- Investigating ways to leverage interpretability findings to improve model robustness, alignment, and performance.

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

Link Feb 2023 – Sep 2023

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.
- Analyzed 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.

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

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

Teaching Experience

Relevant to Academic Major

- **Head Teaching Assistant: *Generative Models*** *(M.S.)* — Sep 2025 – Present
  Dr. Fatemeh Seyyedsalehi, Sharif University of Technology
- **Teaching Assistant: *Machine Learning Operations*** *(B.S.)* — Sep 2025 – Present
  Dr. Fatemeh Seyyedsalehi, Sharif University of Technology
- **Head Teaching Assistant: *Machine Learning*** *(B.S.)* — Sep 2025 – Present
  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.S. Math & CS)* — Jan 2023 – Jun 2023
  Dr. Fatemeh Mansoori, University of Isfahan
- **Head Teaching Assistant: *Operating Systems*** *(B.S.)* — Jan 2023 – Jun 2023
  Dr. Mojtaba Rafiee, University of Isfahan
- **Instructor: *Algorithms and Data Structures*** *(B.S.)* — 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.S.) — 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.S.) — Feb 2025 – Jun 2025
  Dr. Fatemeh Seyyedsalehi, Sharif University of Technology — Grade: 20.00 / 20.00
  PY GPT2 | PY Math Reasoning in LLMs | PY PINN | PY DQL Atari | PY SSL & Contrast. Learning | PY FNN from Scratch

Generative Models (M.S.)

Sep 2024 – Feb 2025

Dr. Fatemeh Seyyedsalehi, Sharif University of Technology

Grade: 18.00 / 20.00

- PY RNN & Transformer
- PY Diffusion/DDPM
- PY EBMs
- PY β-VAE
- PY Pix2Pix
- PY Norm. Flows
- PY Bayesian Network

Machine Learning (M.S.)

Sep 2024 – Feb 2025

Dr. Ali Sharifi Zarchi, Sharif University of Technology

Grade: 20.00 / 20.00

- PY BERT & LoRA
- PY Word2Vec
- PY Knowl. Distil. w CLIP
- PY MobileNet, Eff. & Knowl. Distil.
- PY VAE Img Coloring

Matrix Computations (M.S.)

Sep 2023 – Feb 2024

Dr. Mohammad Reza Razvan, Sharif University of Technology

Grade: 19.00 / 20.00

- PY SVD RGB-Image Compressor from Scratch
- MATLAB Comparing Linear Solvers Implemented from Scratch

Computer Networking (B.S.)

Feb 2023 – Jun 2023

Dr. Mohsen Alambardar, University of Isfahan

Grade: 18.75 / 20.00

- PY Group Chat Application with TCP Protocol and Highspeed Data Sharing

Artificial Intelligence (B.S.)

Sep 2022 – Feb 2023

Dr. Fatemeh Mansoori, University of Isfahan

Grade: 19.50 / 20.00

- PY AI Search Algorithms for Solving Mazes
- PY MDP Gridworld Solver with Value & Policy Iteration

Cryptography (B.S.)

Sep 2022 – Feb 2023

Dr. Mojtaba Rafiee, University of Isfahan

Grade: 19.50 / 20.00

- Feasibility of Proving Authenticity and Data Integrity for P2P Using Hash

Fundamentals of Operating Systems (B.S.)

Feb 2022 – Jun 2022

Dr. Mojtaba Rafiee, University of Isfahan

Grade: 19.50 / 20.00

- PY Reader-Writer Lock Implementation with Three Priorities

Mathematical Databases (B.S.)

Feb 2022 – Jun 2022

Dr. Fatemeh Mansoori, University of Isfahan

Grade: 19.60 / 20.00

- PostgreSQL University Course Registration System
- PostgreSQL Academic Transcript Generator & Analyzer

Mathematical Softwares (B.S.)

Feb 2022 – Jun 2022

Dr. Najmeh Hoseini, University of Isfahan

Grade: 17.75 / 20.00

- MATLAB CLI University Portal for Grade Management

Algorithms & Data Structures (B.S.)

Feb 2022 – Jun 2022

Dr. Jaafar Almasizadeh, University of Isfahan

Grade: 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.S.)

Feb 2020 – Jun 2020

Dr. Jaafar Almasizadeh, University of Isfahan

Grade: 20.00 / 20.00

- PY RSA Cryptosystem Implementation from Scratch
- PY CLI Banking System

Selected Presentations

- **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
- **Universal and Transferable Adversarial Attacks on Aligned LMs** Link May 2025
  Deep Learning Seminars, Sharif University of Technology
- **Propagating Universal Perturbations to Attack LLM Guard-Rails** 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	Languages / Core Tools	Modules / Frameworks	Technical Softwares
ML/DL	Python	PyTorch	Illustrator
GMs/LLMs	R	TensorFlow	InDesign
XAI	MATLAB	Scikit-learn	Photoshop
NeSy AI	PostgreSQL	NumPy	Premiere Pro
Reasoning	Git	Pandas	After Effects
RAG	Bash/Shell	Transformers	Audition
MLOps	HTML/CSS	Diffusers	Figma
Bioinform	LaTeX	Threading	Canvas
CogNeuro	Sockets	RISC-V	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 Competetive 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

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

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

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

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

Awards

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

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

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

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

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

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

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

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

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

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

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

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

Languages

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