Born 11.12.1991

rahmani.b91@gmail.com

GitHub LinkedIn

Personal page

See my full CV

Languages

English	★★★
Persian	★★★
French	★☆☆

Expertise

Machine Learning

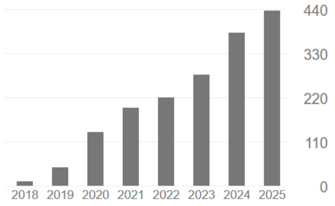
Large Language Models

Diffusion Models

System Identification for Physics/Biology

Publications

Listed on Google Scholar



Community Service

- Assistant Program Chair, NeurIPS 2024
- Co-organizer, NeurIPS 2023–2024 workshop MLNCP

Grants & Awards

- EPFL eSeed 2020, 100K CHF
- Marie Curie Fellowship 2023, €174K, BiTFormer

Tools

PyTorch TensorFlow Git
Python C/C++
AzureML Docker
Large-model training
Prompting API integration
MATLAB Linux Windows

Babak Rahmani

About me

My research focuses on **model architecture** and **efficient training/inference** for foundation models, spanning recurrent/implicit sequence models, agentic scaffolds, and code/physical-AI testbeds. Experience includes large-scale training (7B-class), rigorous evaluation/robustness, and end-to-end research engineering (data/trace pipelines and systems-aware implementation).

Education

2018 – 2022 **PhD Electrical Engineering (EE)** EPFL, Lausanne, Switzerland
Thesis: Learning of physical systems: from inference to control
Supervisors: Christopher Moser & Demetri Psaltis

2014 – 2016 **MSc EE** Sharif University of Technology, Tehran, Iran
GPA: 17.77/20.00 (3.79/4)

2010 – 2014 **BSc EE** Tehran University, Tehran, Iran
Ranked top 10/120. GPA: 18.03/20.00 (3.88/4)

Experience

2025 – now **Visiting Researcher (Sabbatical), Tübingen ELLIS & AI Center**
Marie Skłodowska-Curie Fellow (BiTFormer). Research on **agentic systems, world models, and open-ended reasoning**; building agentic scaffolds for open-ended tasks via inference-time adaptation and RL.

2022 – now **Researcher, Microsoft Research (Cambridge, UK)**

- **Code intelligence & world models**: improving LLMs for code generation and verification under the umbrella of Code World Models (CWMs). Identified fundamental issues in current code-world-modeling LLMs around long-horizon code execution state tracking and efficiency, and developed a linear-RNN approach as a solution. **Learning State-Tracking from Code Using Linear RNNs** (2026, co-first author; supervisor) and **Debugging Code World Models** (2026).
- **LLM architecture & efficiency**: built large-scale recurrent language models (Recurrent LLaMA and Recurrent Mamba) trained on 200B+ tokens; observed stronger reasoning per parameter than standard transformers, at the cost of higher FLOPs/token. Improved efficiency by parallelizing recurrency and retrofitting standard pretrained models into recurrent ones to reduce pretraining cost. **ICML 2025 (Spotlight)**: **Implicit Language Models are RNNs: Balancing Parallelization and Expressivity**. Implicit/recurrent computation improves robustness and generalization beyond language (**Regularizing the Infinite**).
- **New compute & physical AI**: Co-led the ML effort (15+ person collaboration) on the Analog Optical Computer (AOC), an analog-optical compute stack for energy-efficient inference and combinatorial optimization; translated hardware constraints into model abstractions and evaluated generalization/robustness (**Nature**). Developed algorithms for training physical neural networks: backprop-free local learning (**Science**) and efficient training mechanisms (**Nature**), requiring a trinity of software–system–hardware co-design.

2018 – 2022 **PhD Student, EPFL, Switzerland**

- **Biology / neural control**: probabilistic modeling + control of retinal ganglion cell spiking in mice. **NeurIPS**; validated in retina samples (**Nature Communications**).
- **Physics / system identification**: learning-based identification and control of nonlinear time-varying optical systems (**Nature Machine Intelligence**).