Erfan Miahi — ML Research Engineer

Canada – Toronto

✉ mhi.erfan1@gmail.com ▪ (+1) 8254360215 👤 erfanmiahi.com

HIGHLIGHT OF SKILLS & QUALIFICATIONS

- **3+ years in MLE/SWE**: Built scalable Machine Learning (ML) pipelines, delivering impactful real-world solutions.
- **8+ years in AI Research**: Published **6 top-tier papers** with **100+ citations**, collaborating with **Google DeepMind**.
- **End-to-end understanding of LLMs**: Built GPU kernels in Triton, distributed LLM training systems, and AI agents.
- **Exceptional communication skills**: Excel at mentoring diverse individuals across technical and business domains.
- **Proven project management**: Successfully managed multiple projects while maintaining quality within timelines.

RELEVANT EXPERIENCES

Machine Learning Research Engineer — **November 2022 - October 2024 (2 years)**
DeepR Analytics — *Remote, Canada & US*

Part of the founding ML team developing a novel large-scale ML system for automated trading, where I:

- **Developed ML pipeline and algorithms in PyTorch**, driving **$35,000+ daily profit increase per asset**.
- **Migrated pipeline to Amazon ECS**, enabling **large-scale deployment and experimentation** on various trading assets.
- **Optimized ML Pipeline for 9x faster model training**, reducing resource usage significantly.
- **Achieved 85% test coverage with Pytest, reducing bugs by 45%** across ECS, GitHub CI, and C++ Trading Engine.
- **Implemented and analyzed 15+ ML papers** in PyTorch, boosting **profit by $15,000+ per asset**.

CEO — **August 2023 - August 2024 (1 year)**
*IntractAI (**Leadership and Business**)* — *Toronto, Canada*

- **Led a team of 2 engineers/researchers**, driving the development of backend infrastructure.
- Built backend infrastructure for **continual LLM fine-tuning** on the codebase and internal documents.
- Developed a **RAG pipeline** and **web-scraping agent** to improve code generation by **20%** on an internal benchmark.
- Optimized LLM models for **4x faster inference** using quantization, VLLM library, and flash-attention.

Research Associate — **Sep 2022 - Dec 2022 (4 months)**
University of Alberta (RLAI Lab) — *Edmonton, Canada*

- **Redesigned ML codebase** for sensory data prediction, accelerating experimentation and analysis by **50%** (Publication).
- **Developed and deployed an optimized control algorithm**, improving system performance by **21%**.

Assistant AI Consultant — **Nov 2021 - Mar 2023 (1 year 5 months)**
Alberta Machine Intelligence Institute (Amii) — *Edmonton, Canada*

- Provided continuous technical coaching to 5 enterprise teams, aligning projects with their business needs.
- Consulted an energy firm on problem formulation and ML design, **potentially saving $10-15M annually** (Blog).

AI Mentor (For More Info) — **Jan 2017 - present (7+ years)**
*Self-employed (**Leadership**)* — *Remote/Edmonton, Canada*

- Mentoring **10+ students** (past and present) in AI research and engineering for **7+ years**.
- Mentored a University of Guilan (rank 1422) student to secure a PhD at **Michigan State** University (rank 116).

Software Developer — **Oct 2016 – Nov 2017 (1 year 2 months)**
Nilasoft — *Rasht, Iran*

- **Built and optimized** a Spring/Angular e-commerce platform for **1K+ users** with **99.9% uptime**.

EDUCATION

M.Sc. in Computing Science — **Jan 2021 - Aug 2022**
University of Alberta — *Edmonton, Canada*

- **Supervisors**: *Martha White* & *Marlos C. Machado*
- **Thesis**: Feature Generalization in Deep RL: An Investigation into Representation Properties
- **Committee**: *Martha White* (**UoA**) & *Marlos C. Machado* (**Ex DeepMind**) & *Matthew E. Taylor* (**Ex BorealisAI**)

B.Sc in Computer Engineering — **Sep 2015 - Aug 2019**
University of Guilan — *Rasht, Iran*

- Ranked 1st in class

RELEVANT PROJECTS (*60+ GitHub Projects*)

Llamma Circuit Analyser: Converting Sketches to Interactive Circuits in Meta Quest 3 — Backend / Frontend

LLM-Powered Contract Review and Extraction for Revenue Accountants Repo
AI-first Software Development Copilot for Specialized Industries (e.g., Optics) Backend
Mixed-Precision Library for LLM Training with Torch JIT and Bitwise Manipulation (FP8 and BF16)
Distributed LLM Training Infrastructure using Kubernetes and Terraform (DDP/FSDP in PyTorch)
Distributed RL Framework: A3C/PPO/IMPALA Implementations Repo
Open-Source RL Framework for Fast Experimentation on Real-World Problems (Sole Contributor) Repo
Solutions to Assignments of Deep RL Course by UoC, Berkeley (100+ Stars on GitHub) TensorFlow/PyTorch

KEY SKILLS

ML Frameworks	**PyTorch** (**Distributed**), TensorFlow, Ray, VLLM, Accelerate, LangChain, Scikit-learn, Optuna
Programming	**Python**, **Triton** (**GPU Kernels**), C++, Java, JavaScript, TypeScript, SQL, HTML, CSS
Infrastructure	Kubernetes, Terraform, Slurm, Docker
Tools & Databases	GitHub Actions, WandB, Git, Vim, FastAPI, PostgreSQL, Redis, ROS
Cloud Services	Google Cloud, AWS Cloud, Compute Canada, Azure, Nebius
Foundations	ML and Deep Learning, RL, Mathematics, Data Structures and Algorithms
Methodologies	Agile, MLOps, CI/CD, TDD, OOP, Design Patterns

SELECTED PUBLICATIONS (*100+ Citations on Google Scholar*)

Investigating the Properties of Neural Network Representations in Reinforcement Learning
 Han Wang, Erfan Miahi, Martha White, Marlos C. Machado, et al. (**DeepMind Collaboration**) Paper
Published in Artificial Intelligence Journal (AIJ)

GVFs in the Real World: Making Predictions Online for Water Treatment
 Muhammad K. Janjua, Haseeb Shah, Martha White, Erfan Miahi, et al. Paper
Published in Machine Learning Journal (MLJ)

Scalable Transfer Evolutionary Optimization: Coping with Big Task Instances
 Mojtaba Shakeri, Erfan Miahi, Abhishek Gupta, Yew-Soon Ong (**NTU & A*STAR Collaboration**) Paper/Code
Published in IEEE Transactions on Cybernetics Journal

Resmax: An Alternative Soft-Greedy Operator for Reinforcement Learning
 Erfan Miahi, Revan MacQueen, Alex Ayoub, Abbas Masoumzadeh, Martha White Paper
Published in Transactions in Machine Learning Journal (TMLR)

Genetic Neural Architecture Search for Automatic Assessment of Human Sperm Images
 Miahi, Erfan, Seyed Abolghasem Mirroshandel, and Alexis Nasr Paper
Published in Expert Systems and Applications

AWARDS AND ACHIEVEMENTS

- **1st Place at Llama Meta Hackathon Toronto 2024 (45 teams and 150 attendees) - Won $25K, Unitree Robot, and 4 Ray-Bans**
- **Reached Y Combinator Summer 2024 interview stage (top 7% applicant)**
- **Placed among the top 10% of reviewers in ICML 2022**
- **Ranked 6th in World Robot Challenge 2018 (received travel grant to Japan)**
- Abrishamchian Reward (Awarded for graduating with the highest distinction among all students of the University)
- Top Researcher in Computer Engineering Group
- Honorable Mention in ACM/ICPC 2016

LANGUAGES

English - Fluent   **Persian** - Native   **French** - Novice