

# Arman Hosseinmardi

## AI/ML Engineer | Software Engineer Candidate

Montreal, Quebec, Canada • armanhosseinmardi@gmail.com • 438-439-3515

• [www.linkedin.com/in/arman-hosseinmardi](https://www.linkedin.com/in/arman-hosseinmardi)

Results-driven AI/ML Engineer with experience building full-stack Generative AI applications and backend systems using Python and cloud-native tools. Proven ability to design modular LLM pipelines (LangChain, RAG, PEFT), integrate with enterprise platforms, and guide cross-team AI adoption through scalable tooling. Passionate about enabling organizations to operationalize GenAI and accelerate innovation.

### **PROFESSIONAL EXPERIENCE**

#### **AI Software Engineer Intern | CREO Solutions | Montreal, Quebec, Canada**

*Aug 2024–Present*

- Architected scalable backend modules for deploying LLM-based NLP solutions using LangChain and Retrieval-Augmented Generation (RAG) architectures, enhancing contextual document querying within enterprise platforms.
- Designed modular AI pipelines to support LLM fine-tuning workflows using private domain-specific corpora, integrated with internal evaluation suites.
- Enabled human-in-the-loop review by integrating semantic validation layers with retrievers, increasing model trustworthiness in enterprise document workflows.
- Developed Generative AI applications using models such as GPT, LLaMA, or Falcon, including embeddings, prompt templating, and semantic search pipelines.

#### **AI Software Engineer and Research Assistant | Concordia University Innovation Lab | Montreal, Quebec, Canada**

*Jan 2024–Present*

- Architected scalable backend modules for deploying LLM-based NLP solutions using **LangChain** and **Retrieval-Augmented Generation (RAG)** architectures, enhancing contextual document querying within enterprise platforms.
- Automated experimentation with parameter-efficient fine-tuning (PEFT, LoRA) for large-scale language models, optimizing performance on resource-constrained GPUs.
- Built scalable pipelines for structured extraction from unstructured documents, including transformer-based OCR post-processing and embedding generation for retrieval.
- Delivered AI workshops and office hours to mentor teams on LangChain, prompt engineering, and evaluation techniques.

#### **Data Science Researcher | Data Science Laboratory, Shahid Beheshti University | Tehran, Iran**

*Jun 2021–July 2023*

- Engineered backend and API infrastructure supporting a large-scale cyberviolence detection system, deploying LLMs and fine-tuned classifiers in cloud-native environments (Azure Functions, Container Apps).
- Conducted research and development of RAG pipelines combining vector search, LangChain toolchains, and domain-specific context assembly for sensitive content detection.
- Created multi-stage NLP workflows integrating model chaining, prompt engineering, and LLM prompt evaluation frameworks (e.g., LMSys).
- Designed modular AI pipelines to support LLM fine-tuning workflows using private domain-specific corpora, integrated with internal evaluation suites.

#### **Teaching Assistant - Fundamentals of Computational Intelligence | Shahid Beheshti University | Tehran, Iran**

*Sep 2021–February 2022*

- Supported delivery and practical application of machine learning concepts, from neural network design to evaluation and optimization.
- Guided students through complex topics including algorithmic bias, performance benchmarking, and hands-on coding in Python.
- Organized supplementary labs and workshops, improving students' skills in NLP and data-driven feature engineering.

**Teaching Assistant - Discrete Mathematics | Shahid Beheshti University | Tehran, Iran**

*Oct 2020–March 2021*

- Co-developed practice materials and solution sets designed to strengthen algorithmic thinking for backend and AI system challenges.
- Clarified abstract concepts and demonstrated practical applications in software logic and system engineering.
- Enabled effective skill transfer, preparing students for technical careers involving large-scale data and AI workflows.

**Teaching Assistant - Advanced Programming | Shahid Beheshti University | Tehran, Iran**

*Nov 2019–June 2021*

- Mentored students in mastering multithreaded backend architectures, data structures, and algorithmic design.
- Developed coding exercises that bridged theory with production-grade software requirements.
- Led weekly coding sessions to enhance comprehension of enterprise-level programming patterns.
- Provided critical code reviews and one-on-one support, laying groundwork for scalable and maintainable software development.
- Integrated model experimentation with MLflow for tracking, reproducibility, and pipeline automation.

**EDUCATION**

**Concordia University | Montreal, Quebec, Canada**

*September 2023 – August 2025*

Master of Applied Science, Information and Quality System Engineering

**Shahid Beheshti University | Tehran, Iran**

*September 2018 - January 2023*

Bachelor's degree, Computer Engineering

**SKILLS**

|                                   |                                |                             |
|-----------------------------------|--------------------------------|-----------------------------|
| Python                            | NoSQL (Elasticsearch, MongoDB) | Data Visualization          |
| Backend Development               | Azure Cloud Services           | Machine Learning Algorithms |
| AI Model Integration              | Full Stack Development         | Lang Chain                  |
| Natural Language Processing (NLP) | CI/CD                          | Large Language Models       |
| MLOps/ LLMOps                     | AI-Human Workflows             | RAG                         |
| Airflow                           | Snowflake                      | PEFT                        |
|                                   | Databricks                     | Spark                       |

**Projects**

- **Enterprise Document RAG Chatbot:** Designed and deployed a LangChain + Azure-based RAG system with semantic validation layers for legal document analysis.
- **GenAI Knowledge Assistant:** Built a Streamlit-based LLM app integrating vector search (FAISS) and fine-tuned LoRA model for domain Q&A.
- **Built full-stack GenAI applications with Flask/FastAPI** (backend) and minimal frontends (Streamlit) for demo and POC delivery.