Amirreza Naziri

Amir79Naziri | amirreza-naziri | website | +1(437)818-8378
naziriamirreza@gmail.com

Summary

AI / Machine Learning Engineer with **3+ years of applied ML experience**, focused on **Generative AI, LLM-based systems, and cloud-deployed AI services**. Experienced in building **REST-based APIs, and production-oriented ML pipelines on AWS**, translating research ideas into scalable, business-facing AI solutions across NLP and complex data domains.

Work Experience

Machine Learning Researcher, York University Jan 2024 – Dec 2025

- Research supervised by **Vector Institute affiliated faculty (Dr. Laleh Seyyed-Kalantari)**.
- Developed and fine-tuned **Generative Models** (Diffusion, VAEs) and **LLMs** for predictive modeling across **1TB+** high-dimensional datasets (**19K features**).
- Applied advanced architectures including **Graph Neural Networks (GNNs)** to large foundation models (scFoundation, scGPT), improving accuracy by **90%+** and accelerating inference workflows.
- Engineered and prototyped **scalable ML reference applications and POCs**, translating research outputs into **production-oriented solutions** suitable for **API-based deployment**.
- Collaborated with engineers to package models for **deployment-ready environments**, emphasizing reproducibility, scalability, and monitoring.

Machine Learning Engineer, Sharif DeepMine Jul 2021 – Sep 2021

- Developed and optimized **NLP pipelines for production workloads**, processing **10K+ documents/hour** using Python, NLTK, and SpaCy.
- Designed preprocessing components supporting **real-time and batch inference systems** for large-scale text analytics.
- Evaluated open-source NLP tools to recommend an optimal technology stack for scalable data pipelines.

Machine Learning Instructor and Mentor, AI4Good Lab (Mila) May 2025 – Jun 2025

- Mentored a two-month cohort of 5 students, guiding teams to design and deliver **applied ML and GenAI solutions** for social-impact use cases.
- Delivered 20+ hours of lectures and live-coding workshops covering **generative models, deployment considerations, and responsible AI**.

Machine Learning Researcher, Amirkabir University of Technology Sep 2022 – Sep 2023

- Fine-tuned **BERT-based models** for NLP automation tasks, improving text accuracy by **95%+**.
- Developed and deployed a **full-stack Django application** exposing model inference through a user-facing interface.
- Contributed to open-source NLP libraries, extending support for underrepresented languages.

Education

2024 – Jan 2026	M.Sc. Computer Science, **York University**	(GPA: **4.0/4.0**)
2018 – 2023	B.Sc. Computer Engineering, **Amirkabir University of Technology**	(GPA: 18.94/20)

Skills

LLM & Generative AI	PyTorch, TensorFlow, HuggingFace, **LLM Fine-tuning**, **RAG Pipelines**, **Agentic AI**, MCP, ADK, LangChain, Transformers, Diffusion Models, VAEs, GNNs
Cloud, APIs & MLOps	**AWS (S3, EC2, Redshift)**, Docker, Kubernetes, **FastAPI (REST APIs)**, CI/CD, Model Monitoring (W&B), Django, Flask, Linux, Bash
Data Platforms	Spark (PySpark, SparkML), Hadoop (MapReduce), Hive, SLURM
Programming	**Python**, JavaScript (Node.js), SQL, R
Databases	PostgreSQL, MySQL, MongoDB

Publications

Asgari, Arash et al. (Nov. 2025). "MedPerturbing LLMs: A Comparative Study of Toxicity, Prompt Tuning, and Jailbreaks in Medical QA". In: *Proceedings of the AAAI Symposium Series* 7.1, pp. 438–447. DOI: `10.1609/aaaiss.v7i1.36916`. URL: `https://ojs.aaai.org/index.php/AAAI-SS/article/view/36916`.

Naziri, Amirreza, Arash Asgari, Aijun An, et al. (2025). "From Bias to Breakdown: Benchmarking Failure Mode Analysis of Single-cell RNA Sequencing Foundation Models in Acute Myeloid Leukemia". In: *Proceedings of the AAAI Symposium Series*. Vol. 7. 1, pp. 553–557.

Naziri, Amirreza, Arash Asgari, Eleftherios Sachlos, et al. (2025). "Improving Classification of Cell Types in Acute Myeloid Leukemia with Self-guided Masking Technique". In: *NeurIPS 2025 Workshop on AI Virtual Cells and Instruments: A New Era in Drug Discovery and Development*. NeurIPS, AI4D3.

Projects

Spell Correction App using BERT Transformer GitHub

Django-based application for Persian spell correction using BERT and edit-distance methods, designed as a **REST-based AI service** suitable for external system integration.

Real-Time Twitter Sentiment Analysis Pipeline GitHub

Built a **real-time, low-latency inference pipeline** using Spark Streaming and Spark ML, classifying **10,000 tweets/sec** with sub-**50ms latency**.

Additional Projects GitHub Portfolio

Additional projects include **LLM fine-tuning,**, diffusion modeling, and applied biomedical ML.

Volunteering

Trainee Representative, Knowledge Mobilization, Connected Minds Oct 2024 – Sep 2025

- Represented trainee researchers on a cross-functional steering committee, translating technical AI research into actionable insights for non-technical stakeholders.

Last updated: December 24, 2025