Shirin Shahabi

365-883-4747 | Toronto, ON | Gmail | linkedin.com/in/shirin-shahabinejad | github.com/shirin-shahabi| Personal Website

Profile

Machine Learning Engineer with 4 years of hands-on experience in the Finance and Tech industries with an MBA (Finance) and MSc in Computer Science. Proven track record in delivering end-to-end GenAI solutions, driving data-driven decision-making, and fostering cross-functional collaboration. Expert in **Verifiable AI Inference, Reinforcement Learning, Agentic AI and Compression in GenAI Solutions**. Proficient in advanced techniques including Retrieval-Augmented Generation (RAG) systems, Reinforcement Learning with Human Feedback (RLHF), Group Relative Policy Optimization (GRPO), Multi-Context Protocol (MCP) development, and Supervised Fine-Tuning (SFT).

Technical expertise in **AWS (Lambda, EC2, SageMaker) with Cloud Practitioner certification, and GCP (Vertex AI, Kubernetes, Docker, VM instances)**.

Experience

Machine Learning Engineer — May 2024 – Present
Inference Labs Inc (Verifiable AI-Inference) — Hamilton, Canada

- Led the development of verifiable LLM on Qwen 2.5 with no overhead of computation and designed Automated testing using **Multi Context Protocol(MCP) sampling** .
- Led the Quantization and post-training team and Infrastructure on Open-source LLM Models.
- Developed the first Verifiable Agentic AI with proof of inference, enabling verifiable model predictions to **meet legal and insurance liability requirements: LinkedIn Post.**
- Engineered a comprehensive **automated SR&ED auditing documentation pipeline** using Google's Flan T5, seamlessly integrating GitHub, Slack messages, and Jira to streamline employee documentation.
- Designed and implemented a **ranking system** based on DeepFM and BERT, utilizing Reinforcement Learning with Human Feedback (**RLHF**) and Direct Preference Optimization (**DPO**), containerized with **Docker**.
- Developed **Autonomous Index Agent with optimizing liquidity and slippage.**
- Improved inference time by 23% by developing a Commit-Reveal strategy on **Kubernetes using GCP instances.**
- Developed Cosmos SDK modules for decentralized AI inference platforms.
- Developed a Retrieval-Augmented Generation (RAG) system using LLaMA 3 for the R&D team specializing in Zero-Knowledge Cryptography deployed on **Graph VertexAI.**
- Developed Verifiable Ray-tracing physical simulation based on Nividia Warp , Google deepmind Mujoco and Autodesk XLB frameworks.
- Applied structured pruning and quantization techniques on time series prediction models using **NVIDIA's pruning methods** for 8-billion parameter LLaMA and ResNet models without loss of accuracy.

Machine Learning Developer — Dec 2022 – Sep 2024
McMaster University — Hamilton, Canada

- **Developed an Trading Bot** , with enhanced risk mitigation strategies to prevent market crashes, leveraging FastDQN reinforcement learning to outperform human traders by 185%.
- **Implemented multi-objective bi-strategy West America's rail-highway transportation optimization on high-performance computing cluster using Groubi multi-threading and shell scripting** focusing on class based **Supply Chain Managemen.**
- **Implemented a multi-objective bi-strategy optimization model using hierarchical reinforcement learning**, improving conventional Multi-Armed Bandit problems by 13% with **Meta's PEARL framework.**
- **Instructed "Statistical Data Analysis in Healthcare"** using Python (EHealth 705), with materials available at: github.com/shirin-shahabi/PythonTutorial_Ehealth705.

Data Scientist — Feb 2022 – Jan 2023
Nobitex — Remote

- Developed a scalable, robust, and **interpretable ensemble tree-based algorithm** for clustering over 4 million customers by leveraging Spark for distributed transaction data processing, resulting in increasing MAU from 4M to 8.5M in under a year.
- Led cross-team collaboration with the Product Team to design an **end-to-end Marketing Data Modeling and ETL Pipeline using Django and ORM Design.**
- Defined +50 financial KPIs and identified company priorities and targeted measures for strategic growth.
- Utilized Web Engage for Customer Journey Mapping, achieving a 35% increase in conversion rate.

- Conducted market research and customer survey analysis, driving a 10% increase in user engagement.
- Collaborated on Power BI and EDA Course for Managers, promoting dashboard adoption for data-driven decision-making.
- Conducted Instagram and Twitter sentiment analysis, extracting key insights to inform marketing strategies, resulting in a 30% increase in retention rate.

Financial Analyst   Jan 2022 – Feb 2022
Tourism Bank   Tehran, Iran
- **Designed a comprehensive business plan** for a local +12 million dollar hotel construction project.
- Conducted cash flow projections and performed a feasibility study.

Data Analyst   Oct 2020 – Dec 2021
Rocket Internet (SnappMarket)   Remote, Canada
- Contributed to a visionary product with over $100 million in funding for a **fully automated shopping experience** (Low-Cost Amazon Go), emphasizing indoor location tracking, in collaboration with Rocket Internet.
- Optimized inventory reordering for 15 hypermarkets with a fully automated algorithm for +8,000 SKUs.
- Collaborated on loyalty, **fraud detection**, retention, targeted marketing, and churn prediction programs.
- Mentored two interns in the Data Team, now successful Data Scientists one pursuing phd in University of Michigan and the other doing master of computer science at Waterloo University.
- Delivered +100 optimized SQL queries, benefiting Marketing, Operations, and Finance teams.
- Established real-time operational KPIs with automated email reports for timely insights, ensuring data accuracy.
- Defined packaging strategies and presented enhanced user experiences, resulting in a direct 10% increase in revenue.

EDUCATION

M.Sc. Computer Science   Dec 2022 - Dec 2024
McMaster University; GPA: 4 / 4 - Full Scholarship and Fellowship funds

Thesis: Multi-Objective Network Weight Optimization through Hierarchical Reinforcement Learning Strategy.

- Member of Degroote finance and investment council (DFIC Club) - Quantitative Finance Department.

- Completed the Vector Institute 2024 Deep Learning and Reinforcement Learning Summer School.

MBA, Finance   Oct 2021 - Dec 2022
Sharif University; GPA: 3.8 / 4

Awards: Merit-based Admission by the Office of Exceptional Talents.

B.Sc. Industrial Engineering   Oct 2017 - Oct 2021
Sharif University; GPA: 3.7 / 4

Awards: Member of National Elite Foundation (INEF).

SKILLS

Programming Languages: Python, R, SQL (MySQL, Postgres, MongoDB)
Cloud Platforms: Google Cloud Platform (Kubernetes, Vertex AI Engine), AWS(Lambda, EC2, Sagemaker)

Machine Learning Frameworks: TensorFlow, PyTorch, Scikit-learn, Hugging Face

AI and NLP Tools: LangChain, NLTK, SpaCy, ONNX

MLOps and DevOps Tools: Docker, Git, CI/CD (Argo workflow, Redis, Spark)

Web and API Technologies: REST APIs, Streamlit, Shiny App

Data Engineering: ETL Pipeline Development, Data Modeling, Data Streaming, Distributed Computing
Expertise: Reinforcement Learning, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Natural Language Processing (NLP), Inference Engineering, Compression Techniques, Pre-training and Fine-Tuning, Predictive Modeling, Statistical Analysis, Fraud detection, Verifiable Inference, Agentic AI

Operation Research: Mathematical Optimization, Integer Programming, Stochastic Modeling, Discrete-event Simulation, Meta-heuristic search Algorithms, Process Mining