Faraz Khadivpour

Vancouver, BC | +1-438-979-8252 | khadivpour@sisaenergy.com | farazkhadivpour | fkhadivpour

Summary

AI and ML Engineer with 6+ years of experience designing, deploying and operating production machine learning and generative AI systems. Currently own an **LLM microservice** serving three products end to end, covering **LLM API integration**, **RAG** and vector search, **agentic workflows**, data pipelines, backend APIs, and the evaluation and monitoring around all of it. Background spans deep learning research (M.Sc. Computing Science, Amii) and enterprise data science at Scotiabank, with published work in explainable AI. Owns delivery across the full lifecycle, from requirements and architecture through implementation, code review, deployment and production support.

Work Experience

• AI Engineer — *Jan 2025 – Present*

SISA Energy — Vancouver, BC

- Own and operate **SISA AI**, a production LLM microservice consumed by three products (**InPort**, **CollectFlow**, **VSM**) through a versioned API, centralizing model selection, prompting and provider routing across **Anthropic**, **OpenAI** and **Google** via **LiteLLM**.
- Lead end-to-end design and delivery of production **GenAI** systems that extract validated, structured data from unstructured sources (PDFs, images, CSVs) and internal data stores, automating processing across tens of thousands of buildings.
- Designed and scaled modular **RAG** and **vector search** pipelines with **LangChain** and **LangSmith**, including chunking and retrieval strategy and structured output validation.
- Built an **agent layer** on the **Model Context Protocol (MCP)**, enabling natural language interfaces over platform data through governed, permission-aware access.
- Developed an equipment image understanding pipeline that reduced manual processing time per building from 2–3 days to under 1 hour at **95%+ accuracy** on noisy, low-resolution data.
- Designed and deployed backend services via a production **Flask API**, including request validation, structured outputs and integrations with **PostgreSQL/Supabase** and **Redis** caching for performance and reliability.
- Own evaluation end to end: dataset preparation, offline benchmarking, and online quality monitoring, with custom metrics for accuracy, reliability, latency and cost, and prompt templates managed as versioned code.
- Implemented observability and incident response with **Sentry** and **PagerDuty**, including monitoring dashboards and on-call workflows supporting production SLAs.
- Product owner and core developer for **InPort** and **CollectFlow**, owning delivery from requirements and RFC through architecture, implementation, code review, testing, deployment and production support.
- Use **AI-assisted development** (Claude Code, Cursor) as a core part of the workflow, with specification-driven planning, small pull requests and human review before merge.
- Collaborate cross-functionally in an agile environment with **Git**, **Jira** and **Confluence**, working directly with leadership and clients to align technical decisions with business priorities.

• Machine Learning Researcher — *Jan 2023 - Jan 2025*

Alberta Machine Intelligence Institute (Amii) — Edmonton, AB

- Applied a novel explainable AI method to **Large Language Models (GPT-2)** and **Diffusion Models**, working with transformer internals in **PyTorch**.
- Implemented an approach to improve human understanding of GPT-2 fine-tuned with **LoRA** (Low-Rank Adaptation).

• Data Scientist — *Jan 2022 - Jan 2023*

Scotiabank — Edmonton, AB

- Developed a neural network combining **LSTM** and **attention layers** over 800,000+ loan applications and credit report records in a regulated enterprise environment.
- Worked with large-scale enterprise data in **Python** and **SQL**, covering data preparation, feature processing and model evaluation.
- Collaborated with research partners to apply **explainable AI** methods to the model, improving interpretability metrics by 15%.

• Machine Learning Researcher — *May 2020 - Dec 2021*

Alberta Machine Intelligence Institute (Amii) — Edmonton, AB

- Proposed a novel explainable AI method making neural networks more understandable to human users (Responsibility), published at AIIDE EXAG.
- Built and analyzed models in **Keras** and **TensorFlow** on datasets of 50,000+ images, working with **ResNet**, **AlexNet** and **VGG**.
- Improved interpretability metrics by 5% on image classification and 20% on text classification tasks.
- Designed and ran a **300-participant human evaluation study**, analyzing results in **RStudio** and establishing a 30% preference over baseline approaches.

- **Machine Learning Researcher** *Jan 2020 - May 2020*

  Neuromuscular Control & Biomechanics Laboratory (NCBLab) Edmonton, AB

  - Delivered a client project for RWDI consulting with a team of 5, hitting milestones 2 weeks ahead of schedule.
  - Wrote Python tooling to extract spatial data from 3D building models, preprocessed 10,000+ samples and applied **PCA** and **autoencoders** for a 90% reduction in data complexity.
  - Trained models in **PyTorch**, achieving 94% prediction accuracy for a key variable in wind tunnel simulations.

Education

- **M.Sc. in Computing Science** *Jan 2023 - Dec 2024*

  University of Alberta Edmonton, AB

  GPA: 4.0

- **M.Sc. in Environmental Engineering** , University of Tehran *Sep 2015 - Feb 2018*
- **B.Sc. in Civil Engineering** , K.N.Toosi University of Technology *Sep 2010 - Aug 2015*

Skills

- **Programming:** Python (primary), SQL, JavaScript, R, C#
- **LLM & Generative AI:** LLM APIs (Anthropic, OpenAI, Google, Azure OpenAI), prompt engineering, LangChain, LangSmith, LiteLLM, CrewAI, Braintrust, LLM Guard, Model Context Protocol (MCP), AI agents
- **RAG & Retrieval:** Retrieval-Augmented Generation, vector databases (Chroma, FAISS), semantic search, chunking and retrieval strategy
- **ML & Deep Learning:** PyTorch, TensorFlow, Keras, scikit-learn, transformers, LLMs, CNN, RNN/LSTM, attention, autoencoders, diffusion models, LoRA fine-tuning
- **MLOps & Evaluation:** Model deployment and serving, evaluation pipelines, offline benchmarking, online quality monitoring, custom metrics, versioning, model lifecycle management
- **Backend & APIs:** Flask, FastAPI, Celery, React, Streamlit, REST API design, versioned service interfaces, third-party integrations
- **Data Engineering:** Data pipelines, ETL, schema design, data quality debugging, Pandas, NumPy, Matplotlib, Plotly, Seaborn, Power BI
- **Databases & Caching:** PostgreSQL, TimescaleDB, Supabase, SQLModel, Redis
- **Cloud & Deployment:** Microsoft Azure, Google Cloud (Vertex AI), AWS SageMaker, Docker, Kubernetes, Railway
- **Observability:** Sentry, PagerDuty, LangSmith, monitoring dashboards, incident response
- **AI-Assisted Development:** Claude Code, Cursor, specification-driven development, small-PR practice with human review
- **Collaboration:** Git, Jira, Confluence, Miro

Certificates & Awards

- **Alberta Graduate Excellence Scholarship (AGES)** , University of Alberta *Jan 2025*
- **Advanced Manufacturing Engineers Upskilling Program** , Foresight Canada *Winter 2025*
- **MLOps Specialization** , Coursera / deeplearning.ai *Summer 2022*
- **Machine Learning Technician Certification** , Amii *Fall 2020*
- **Deep Learning Specialization** , Coursera / deeplearning.ai *Summer 2020*

Publications

[1] Khadivpour F, Banerjee A, and Guzdial M. "Responsibility: An example-based explainable AI approach via training process inspection." arXiv preprint arXiv:2209.03433v1 (2022).

[2] Khadivpour F, Guzdial M. Explainability via Responsibility. The 2020 Intelligence and Interactive Digital Entertainment (AIIDE) Workshop on Experimental AI in Games (EXAG). DOI: arXiv:2010.01676.

[3] Shayesteh AA, Koohshekan O, Khadivpour F, Kian M, Ghasemzadeh R, Pazoki M. Industrial waste management using the rapid impact assessment matrix method for an industrial park. Global Journal of Environmental Science and Management. 2020 Apr 1;6(2):261-74. DOI: 10.22034/GJESM.2020.02.10

Teaching Experience

- **Graduate Teaching Assistant, Game AI** *Jan 2023 - Apr 2023*

  Computing Science Department, University of Alberta Edmonton, AB

  - Led tutorials and review sessions for 50+ students on AI techniques in game development, graded assessments and provided detailed feedback on **C#** and **Unity** work.