NADIR TRAPSIDA

trapsidanadir@gmail.com | 819 580-9669 | Montreal, QC, Canada
GitHub | LinkedIn | Portfolio | Medium | LinkTree

Summary

AI/ML Engineer with a computer engineering and electronics background, specializing in bringing machine learning models from prototype to production at scale. Track record of shipping end-to-end ML pipelines, optimizing inference systems serving 150K+ daily requests, and building products used by major North American fleets.

Professional Experience

Machine Learning Engineer *Sep 2024 – Present*
Isaac Instruments Montreal, QC

- Reduced AI inference latency by 50% (80 ms → 40 ms) serving 150K+ daily requests by re-architecting model serving infrastructure and optimizing the API layer, cutting average end-to-end request time by 33%.
- Designed and shipped an end-to-end video analytics pipeline processing 7K+ videos/hour through 20 ML models, collaborating with DevOps on integrated monitoring, resource-efficient scheduling, and automated recovery to bring the service from prototype to production.
- Built a TTS/STT voice-command assistant using OpenWakeWord for truck drivers on resource-constrained edge hardware, enabling hands-free interaction to improve driver experience and safety while on the road.
- Identified a critical gap in the existing scanning solution, then partnered with the product team to build a POC and drive it to a production feature that enhanced tablet-based document scanning and data extraction across the entire Isaac client base.
- Developed a "Driver Year in Review" experience with gamification elements (stats, badges, milestones) to boost driver retention and engagement across partner fleets.
- Accelerated dataset creation across multiple teams by building an internal image/video annotation tool that streamlined labeling throughput and review cycles.
- Delivered a safety-event application used by major Canadian and US fleets by engineering the FastAPI backend and contributing to the Next.js frontend, providing clients real-time access to safety insights.
- Prototyped a RAG-based internal knowledge search chatbot with hybrid BM25 + vector retrieval, improving document discovery for engineering and operations teams.

Co-Founder & Chief AI Officer *2022 – 2024*
Queva Inc. Montreal, QC

- Architected and deployed one of the first veterinary-assistance conversational agents powered by RAG, increasing customer engagement by 35% and improving retention across the user base.
- Designed the production deployment pipeline for the RAG system, achieving 99.5% uptime and reducing inference errors in a latency-sensitive consumer application.
- Trained a predictive model on IMU sensor data to classify dog activities at 97.5% accuracy, enabling real-time behavioral insights for pet owners.
- Built the IoT backend integrating database, mobile application, and ML inference services, halving the original integration timeline.
- Managed and optimized multi-service infrastructure on Microsoft Azure, reducing monthly cloud spend by 30% while improving scalability.
- Optimized CI/CD pipelines, cutting deployment time by 3 hours and increasing release reliability for the engineering team.
- Designed the mobile application prototype that led to a successful pitch, securing $250K in investment from VCs.
- Managed a cross-functional team of 6 developers and interns, improving project delivery cadence and team output.

AI Research Engineer *2022*
Vehicle Intelligence Laboratory, Université de Sherbrooke (LIV) Sherbrooke, QC

- Developed a multi-camera 3D simulation of an autonomous bus in CARLA, enabling reproducible testing of perception and planning algorithms.

- Improved the object detection model by 13% in detection accuracy on the CARLA simulator, strengthening safety validation for autonomous driving research.
- Restructured and integrated multiple modules of a large autonomous-vehicle research codebase, accelerating experiment iteration and improving maintainability.
- Collaborated with NovaBus to prepare the integration of simulation-validated algorithms into a physical bus platform.

AI Software Developer *2021*
Centre de Recherche Informatique de Montréal (CRIM) Montreal, QC

- Optimized CI/CD workflows using Makefiles and GitHub Actions, reducing deployment and testing cycle times.
- Increased code coverage by 20% through systematic unit test development, improving codebase reliability.
- Discovered a critical security flaw in a production product, preventing potential data breaches and protecting user privacy.

Data Scientist & Computer Vision Engineer *2020 – 2021*
Decathlon Inc. Montreal, QC

- Built a text classification model at 95% accuracy to distinguish Decathlon-related content from generic sport mentions, enabling scalable brand monitoring across social platforms.
- Optimized ML serving endpoints with Protocol Buffers, reducing prediction latency by 83% (12 s $\rightarrow$ 2 s) and improving throughput for downstream consumers.
- Created an end-to-end image search application (frontend, backend, database) using word-vector embeddings, achieving 90% relevance on social media image retrieval.

Hardware & Software Solutions Integration Specialist, R&D *2018 – 2019*
Kontron Inc. Boisbriand, QC

- Automated network performance and power consumption benchmarks using Python and Bash, reducing testing time by 4 hours per cycle and enabling faster hardware qualification.
- Conducted GPU benchmarks (NVIDIA Tesla P4, T4; AMD Vega 4000) across 5 server families, ensuring driver compatibility and performance baselines for client deployments.

Technical Skills

ML & AI: PyTorch, TensorFlow, Keras, Scikit-Learn, ONNX, OpenCV, MLflow, Hugging Face, LangChain, LlamaIndex, LLMs (GPT, LLaMA), RAG Systems, Prompt Engineering, LoRA Fine-tuning, NumPy

Backend & Data: Python, FastAPI, Flask, gRPC, PostgreSQL, MongoDB, MSSQL, Redis, Pandas, PyTest

Frontend: JavaScript, Vue.js, Next.js

Cloud & Infrastructure: AWS (S3, SageMaker, Lambda, ...), Azure (Container Apps (Kubernetes-managed), IoT Hub, DevOps, Functions, Blob Storage, Web Apps, VMs, ...), OVH, Docker, Grafana

DevOps: Git, GitHub Actions, CI/CD Pipelines, Bash

Languages: French (native), English (fluent)

Education

University of Sherbrooke Sherbrooke, QC
M.Eng. in Computer Engineering, Specialization in Artificial Intelligence Academic Excellence Scholarship
B.Eng. in Computer Engineering, Co-op Program

Selected Projects & Publications

- **DocTrack.io** (Co-Founder): Building a B2B SaaS platform that automates document classification, renaming, and expiry tracking for property management companies using LLM-based email ingestion and extraction pipelines.
- **Deploying PyTorch with FastAPI to Handle 120K+ Requests a Day**: Published in Data Science Collective, detailing production ML serving architecture and optimization strategies.
- **We Switched from TensorFlow to PyTorch and This Happened**: Published in Level Up Coding, sharing real-world lessons from a production ML framework migration.
- **Low-Resource Language Translation LLM**: Fine-tuned an LLM with LoRA to translate English into Zarma/Songhai, a native African language unsupported by Google Translate. Published in Data Science Collective and Towards AI.