Alexandra Tulchinsky

Generative & Agentic AI Engineer | Scalable Backend & Cloud Systems

(613) 981-4848 | alexandra.tulchinsky@gmail.com | LinkedIn| Portfolio | GitHub

Professional Summary

AI/ML software engineer with 2+ years of hands-on experience shipping production-grade Generative AI, Agentic AI, and deep-learning systems across broadcast, telco, and optical-network domains. Proven track record in architecting multi-agent solutions, retrieval-augmented generation (RAG) pipelines, and sovereign-cloud compliant LLM deployments that compress hours-long workflows to seconds and unlock new revenue streams. Recognised with the Nokia Business Impact Award (2023).

Core Competencies

- **Generative & Agentic systems:** multi-agent orchestration, tool-use, safety/guardrails, evals
- **RAG & vector search:** query understanding, hybrid retrieval, reranking, offline/online relevance evals
- **LLM adaptation:** SFT, LoRA/QLoRA, prompt/retrieval design, latency/cost optimization
- **Multimodal retrieval & video understanding:** image/audio/text feature extraction, temporal segmentation
- **ML engineering & MLOps:** packaging, CI/CD for models, model/data versioning, monitoring & drift handling
- **Cloud architecture:** scalable, cost-aware designs on AWS/Azure/GCP; security & networking basics
- **Data engineering:** batch/streaming pipelines, warehousing, orchestration, data quality checks
- **Backend APIs:** high-throughput services, auth, pagination, caching, observability; mentoring & agile delivery

Key Solutions

- **Predictive Optical AI:** Applied neural networks to high-bandwidth optical systems to improve performance metrics and enable proactive insights for telecom hardware.
- **Multimodal Video Retrieval:** Built a natural language–driven search pipeline combining visual, audio, and transcript features to deliver sub-5-second results over decades of broadcast archives.
- **Adaptive Relevance Algorithm:** Designed query interpretation logic that dynamically matches user intent to multimodal features, significantly improving discovery accuracy and relevance.
- **Multi-Agent RAG Automation:** Designed autonomous AI agents for knowledge retrieval, cutting support escalations and halving incorrect responses.
- **Enterprise LLM Search:** Delivered fine-tuned, sovereign-compliant search with major accuracy gains and reduced hallucinations.
- **GenAI Workflow Automation:** Replaced multi-week manual processes with AI-powered pipelines, completing in minutes with fewer errors.

Professional Experience

Ross Video – Software Engineer, AI/ML

Jan 2025 – Present

- Cut broadcast video search from 3+ hours to under 5 seconds by developing a multimodal Natural Language Query (NLQ) pipeline by using open-source LLMs (CLIP, CLIP, SBERT, Whisper, Llama) and vector indexing to

extract searchable features (visual, audio, transcripts, facial identities, sentiment, motion) from both live feeds and decades of archival footage.
- Increased search relevance by 60%, achieving 80% accuracy in MVP testing, by designing a custom algorithm that dynamically interprets user queries to identify relevant video segments based on extracted video features for 100+ broadcast professionals on Ross Video's Streamline platform.
- Streamlined AI-powered product support, reducing customer support escalations by 35%, incorrect responses by 50%, and doubled query response speed through scalable, asynchronous pipelines using a multi-agent Retrieval-Augmented Generation (RAG) architecture with LangChain, OpenAI, and AWS cloud infrastructure.
- Fast-tracked core AI enhancements by mentoring 2+ co-op students who shipped production-ready features, improving user experience and product performance.

Nokia – Software Engineer, AI/ML Intern
Sep 2023 – Aug 2024

- Developed and deployed a scalable enterprise search solution for licensing by fine-tuning a Large Language Model (LLM) using internal enterprise data, enabling interactive querying and summarization features. The solution reduced document search and compilation time from several hours to minutes, achieving a 95% increase in response accuracy and an 80% reduction in hallucinations through prompt engineering, supervised learning, and Low-Rank Adaptation (LoRA).
- Optimized and streamlined revenue allocation across diverse business units using Gen AI. Leveraged Hugging Face transformers library, NLP auto-tagging processes, and MS Power Automate to fully automate the workflow. The manual and error-prone process that took 4-6 weeks to complete now takes 2-5 minutes. In addition, achieved a 30% reduction in errors.
- Awarded the Nokia Business Impact Award for delivering innovative AI-driven solutions that significantly enhanced operational efficiency and accuracy.

Lumentum – Data Scientist Intern
Jan 2023 – Apr 2023

- Improved the speed and bandwidth of the Wavelength Selective Switch (WSS) product lines by predicting the direction of light waves using optical data, reducing processing time by 10%.
- Created and deployed an artificial neural network on Azure that predicts light steering in WSS products, achieving an accuracy of 90%.
- Collected, curated and analyzed networking data to provide insight into potential equipment failures, thereby prolonging equipment health and improving overall equipment lifespan.
- Advised key clients (Ciena, HANA and Fujitsu) on their next version of product releases by collecting their feedback and prioritizing new features & capabilities.

Lumentum – Data Systems Analyst Intern
May 2022 – Aug 2022

- Communicated between the R&D, IT, Data Science and Analytics teams to understand business requirements and desired user experience in terms of capabilities and features.
- Collaborated with external vendors (AnyDesk and ConnectWise) on findings and required improvements in future product releases.
- Leveraged feature-by-feature usage data and end-user survey data to gain insight into consumer needs and predict third-party application use, resulting in 1.5 hours of time savings per user/day by optimized use of the software.

- Developed data-driven solutions by mining and analyzing consumer usage data, reducing internal third-party software costs by $25k.

Selected Projects and Open-Source (GitHub)

- **SkinSafe AI - Eczema Dietary Safety Platform:** Created a full-stack AI solution that turns product photos into instant *safe* vs *avoid* ingredient guidance using multimodal analysis and a curated eczema-trigger database.
- **Dental Clinic Web Application:** Built a cloud-hosted platform that streamlines appointment booking, treatment management, and payments for patients, hygienists, and doctors.
- **Amazon Product Recommendation:** Developed an NLP-driven tool that scrapes, analyzes, and visualizes Amazon reviews to guide purchase decisions with data-backed sentiment insights.
- **YouTube Data Analysis:** Transformed raw channel data into actionable recommendations for boosting audience engagement and revenue growth.

Education & Certifications

- B.Sc. (Hons) Computer Science – University of Ottawa, 2024
- Azure Data Science Associate (DP-100), 2023
- Azure Fundamentals (AZ-900), 2021

Awards & Recognition

- Nokia Business Impact Award (2023)
- University of Ottawa Dean's Honour List (2021-2024)

Technical Stack

Languages: Python, TypeScript, Go, Java, C/C++, SQL
AI/ML & GenAI: PyTorch, TensorFlow, JAX, Keras, scikit-learn, NumPy, Pandas, OpenCV, spaCy, Matplotlib, Seaborn, Hugging Face, Transformers, TRL, PEFT, BitsAndBytes, DeepSpeed, CUDA Toolkit, Sentence-Transformers (SBERT), CLIP/XCLIP, Whisper, Spark MLlib, LlamaIndex, LangChain, LangGraph, AutoGen, CrewAI, Semantic Kernel
Knowledge Graph/Ontology: Neo4j, Neptune
Vector & Search: Weaviate, Pinecone, Chroma, Qdrant, pgvector, FAISS, Elasticsearch/OpenSearch
Data & Streaming (ETL): Apache Spark (PySpark, Streaming), Databricks, Kafka, AWS Kinesis, Airflow, Hadoop, Hive
Warehousing, Lakehouse & SQL/NoSQL Databases: Snowflake, Amazon Redshift, BigQuery, PostgreSQL, MongoDB, Firebase, Parquet, Redis
Cloud & Infra: AWS (Bedrock, SageMaker, EKS, S3, CloudFront, Lambda, DynamoDB), Azure (OpenAI, ML Studio, Cognitive Services, AKS)
DevOps, MLOps & Observability: MCP (Model Context Protocol), Docker, Kubernetes, Terraform, MLflow, Kubeflow, GitHub Actions, Jenkins, CircleCI
Model Serving & Optimization: vLLM, Triton Inference Server, ONNX Runtime, TensorRT
Web, APIs & Servers: FastAPI, Flask, Django, Node.js, Next.js, React.js, GraphQL, REST/RESTful, gRPC, **OpenAPI/Swagger**, Nginx, Apache, Bootstrap, HTML, CSS/SCSS
Tooling, Testing & Protocols: Git, SVN, pytest, JUnit, Selenium, Robot, PyATS; SSH, Telnet, Netconf, TFTP, SCP, Rsync; FFmpeg, PyAV,