Andrii Skliar

MTS (Applied Research) @ Contextual AI

Amsterdam, Netherlands • *Citizenship: Netherlands*

andrew.skliar@gmail.com | github.com/askliar | linkedin.com/in/andriiskliar | Andrii Skliar

Personal Profile

AI Research Engineer with 5+ years of experience building and scaling ML systems, with a focus on efficient inference, model compression, and real-world deployment. Co-inventor on multiple granted patents; contributed to ML research adopted in commercial products. Experienced in bridging foundational research and production-ready solutions in LLMs, NAS, and evaluation infrastructure.

Tools & Platforms: Python, PyTorch, Ray, Kubernetes, Docker, GCP, Git
ML Focus Areas: LLMs, MoEs, NAS, Quantization, Inference Optimization, Synthetic Data Generation

Work Experience

Contextual AI — *San Francisco, CA (Remote)*

Member of Technical Staff (Applied Research) — *Jan 2025 – Present*

- Built and scaled synthetic data pipelines for 10+ enterprise LLM workflows.
- Led end-to-end ML system development, improving model quality by 20% and production reliability to 90%.
- Designed evaluation frameworks for instruction-following, grounding, and technical accuracy.

Qualcomm AI Research — *Amsterdam, Netherlands*

Senior Research Engineer — *Oct 2019 – Jan 2025*

- Led hardware-aware ML research under Prof. Max Welling, contributing to DONNA (AIMET), simulated quantization (17% energy savings), and efficient LLM deployments.
- Built Qualcomm's first hybrid LLM, enabling efficient on-device MoE decoding and product integration.
- Co-inventor on 7 US patents (5 granted, 2 pending) related to model efficiency (filed through Qualcomm).

QUVA Lab — *Amsterdam, Netherlands*

Research Intern — *May 2019 – Oct 2019*

- Master's thesis internship in collaboration with QUVA Lab under the supervision of Maurice Weiler.
- Extended the theory of Hyperbolic Neural Networks to Convolutional and Graph Neural Networks.

Selected Publications

Efficient LLM Inference using Dynamic Input Pruning and Cache-Aware Masking (***Patent Pending***)
Proceedings of Machine Learning and Systems (MLSys), 2025

Mixture of Cache-Conditional Experts for Efficient Mobile Device Inference (***Patent Pending***)
Transactions on Machine Learning Research, 2025

Think Big, Generate Quick: LLM-to-SLM for Fast Autoregressive Decoding (***Patented under Qualcomm IP portfolio***)
Proceedings of the 2nd Workshop on Efficient Systems for Foundation Models (ES-FoMo II) at ICML, 2024

Hyperbolic Convolutional Neural Networks
Arxiv Preprint (first author), 2023

Cyclical pruning for sparse neural networks (***Patented under Qualcomm IP portfolio***)
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022

Revisiting single-gated Mixtures of Experts (***Patented under Qualcomm IP portfolio***)
Proceedings of the 33rd British Machine Vision Conference (BMVC), 2022

Simple and Efficient Architectures for Semantic Segmentation
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022

Simulated Quantization, Real Power Savings (***Patented under Qualcomm IP portfolio***)
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022

Distilling optimal neural networks: Rapid search in diverse spaces (***Patented under Qualcomm IP portfolio***)
Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021

Adding object detection skills to visual dialogue agents
Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018

Education

University of Amsterdam — *Amsterdam, Netherlands*

MSc in Artificial Intelligence — *Sep 2017 - Aug 2019*

- GPA: 8.8/10, Cum Laude
- Honours programme with multiple successful research projects.
- Courses in Theoretical and Applied Machine Learning with main focus on Advanced Machine Learning, Deep Learning, Computer Vision, Reinforcement Learning and Information Theory.