Ilia Badanin

badanin.ilia@gmail.com | nov3o | nov3o.github.io
ilia-badanin | Google Scholar

Education

EPFL — *Lausanne, Switzerland*
Data Science MSc — Sep. 2024 – Jul. 2027

ETH Zürich — *Zürich, Switzerland*
Exchange student in Informatics — Sep. 2025 - Jan. 2027

Bauman Moscow State Technical University — *Moscow, Russia*
*CS BSc with Honours, **GPA: 3.98/4**, highest GPA on a program (out of 110 students)* — Sep. 2018 - Jul. 2023

EPFL — *Lausanne, Switzerland*
Exchange student. Researched at two top-ranked labs: CVLab and MLO Lab — Feb. 2022 - Aug. 2022

Experience

Pleias — *Paris, France (Remote)*
Research Intern — Oct. 2026 - Present

Large-scale synthetic pretraining: generating multimodal corpora for LLM/VLM pretraining (**410k** synthetic VQA). Built a robot-grasping video simulation and rendering pipeline: ∼**40 fps**, ∼**110x** faster than Blender Cycles.

LLM Open Source: Hugging Face *tokenizers*, *transformers* — *Remote*
Open-Source Contributor — Oct. 2026 - Present

PRs to *tokenizers* from my fork toki: a compact parallel BPE trainer (**3.5–19x** faster, **54–61%** less memory, identical merges); constant-memory multithreaded encoding (**1.1–2.1x** faster, up to **95%** less memory, identical ids); thread control. PRs to *transformers*: multithreaded quantization at load time (**3–21x** faster on H100) and a model inspector that runs any architecture from its config at near-zero memory. NF4 grouped-GEMM kernel for MoE experts. lazycomb: Python iterators without evaluation over any composition of combinatorial groups.

ETH Zürich — *Zurich, Switzerland*
Research Assistant — Sep. 2025 - Feb. 2026

Working with Xiaozhe Yao and Ana Klimovic on heterogenous distributed LLM serving (accepted at OSDI) and with Alessandro Stolfo and Mrinmaya Sachan on subliminal learning.

ETH AI Center — *Zurich, Switzerland*
AI Engineer — Mar. 2025 - Aug. 2025

I contributed to building an open-source Swiss LLM. Working with Imanol Schlag I created a human labeling platform that **collected 4,000+ annotations** (demo) to improve automated classification, **analyzed 34 TB** of data, and **built an LLM Arena** (demo) to benchmark checkpoints and external models. The Swiss LLM, Apertus, is **the biggest** European Open-Source LLM, trained from scratch on **1800+** languages, subsequently appearing in **Top 10 of the Hugging Face** Trending Model leaderboard, with **300k** downloads per month across all models. Core contributor to serving platform.

EPFL — *Lausanne, Switzerland*
Research Student — Apr. 2022 - Sep. 2025

- **dlab** (NLP, Aug. 2024 - Sep. 2025): created a benchmark for multilingual LLMs, analyzing how models handle concepts across languages to find factual inconsistencies and hallucinations. arXiv.
- **Machine Learning and Optimization Lab** (Jul. 2022 - Nov. 2023): with Martin Jaggi, developed multiplication-free neural networks using tropical algebra, reducing energy consumption by up to **31x** while maintaining high accuracy.
- **Computer Vision Lab** (Apr. 2022 - Nov. 2023): with Soumava Roy and Pascal Fua, improved pose estimation under occlusion with Graph Convolutional Networks, presented at **3DV 2024**.

Publications

CHSM8: A Chess Transformer that Follows Natural-Language Style Prompts — Oct. 2026
Ilia Badanin. *(arXiv preprint, technical report)*

Benchmarking Concept-Spilling Across Languages in LLMs — Jan. 2026
Ilia Badanin, Daniil Dzenhaliou, Imanol Schlag. *(arXiv preprint)*

OpenTela: Unifying Decentralized Computing Resources for Heterogeneous LLM Serving — Dec. 2025
Xiaozhe Yao, Youhe Jiang, **Ilia Badanin**, Qinghao Hu, Robert Matthew Smith, Binhang Yuan, Imanol Schlag, Eiko Yoneki, Ana Klimovic. *(OSDI 2026)*

Democratizing Open and Compliant LLMs for Global Language Environments Sep. 2025
Alejandro Hernández-Cano, ··· **Ilia Badanin**, ··· Imanol Schlag *(ACL 2026)*

Occlusion Resilient 3D Human Pose Estimation Oct. 2023
Soumava Kumar Roy, **Ilia Badanin**, Sina Honari, Pascal Fua, *(International Conference on 3D Vision 2024)*

Projects

- **CHSM8 (Personal Project)** Jul. 2026
  Built CHSM8 solo in 10 days: a 210M-parameter chess LM that follows plain-English style prompts. Pretrained on 70B moves (optimized for throughput 100k→540k moves/s, and MFU 1%→17%), distilled a model, labeled 30M games with a distilled 1B labeler, then SFT + DPO for text conditioning. 1890 Elo; ranks above a few models trained in a more complex setting with search and rewards, and beats every public supervised chess LM. 56-page report (arXiv); open weights, datasets, and recipe.
- **ReskinEngine (Class Project)** Mar. 2026 – Jun. 2026
  Developed a diffusion model-based approach for temporally-consistent video restyling, enabling text-prompted style transfer across game scenes while preserving temporal coherence.
- **Textocal (Personal Project)** Mar. 2025 – Apr. 2025
  Developed an android application and a browser extension that extracts event-related information from text or images and adds it to the user's calendar. Built the app **within 25 hours**. Presented the project to an audience of 50 at Lausanne **AI Tinkerers** meetup.
- **DynaQuant (Class Project)** Feb. 2025 – Mar. 2025
  Was an invited collaborator on the final project for **Stanford's CS234 Reinforcement Learning course**. Contributed to the development of a quantization technique that adapts pre-trained models to downstream datasets, enabling deployment on smaller models while balancing model size and performance trade-offs.
- **FolioLLM (Class project)** May. 2024 - Jun. 2024
  Was an invited collaborator on the final project for **Stanford's CS224N NLP with Deep Learning** course. Led a team of 3 in designing and fine-tuning a 0.5B-parameter LLM to optimize portfolio management and financial domain expertise. Oversaw strategy, solution design, and execution, ensuring high performance across both financial and NLP metrics.

Teaching

Section Leader, Stanford Code in Place 2024, 2026 (Apr. - Jun.)
One of 700 teachers facilitating learning for 10,000+ students in Stanford's extensive online computer science course

Achievements and Awards

Graduated 1st in class, BSc with Honours Jul. 2023
Highest GPA (3.98/4) among 110 students in the CS BSc program at BMSTU

Ironman 70.3 Aug. 2022
Completed Ironman 70.3 challenge along Léman lake. 2 km of swimming, 90 km of bike riding and 21 km of running in 9 hours

Google Hashcode Feb. 2022
Reached **16% worldwide**, top 5th team out of 32 in local hub

Full-ride academic scholarship for bachelor's degree duration Feb. 2022
Awarded for exceptional academic performance at BMSTU, distinguishing me among just **6 recipients from 120 students** in the program

Skills

- **Skills**: PyTorch, Python, C++.
- **Languages**: English (C1), French (B2), German (A1), Russian (Native)