# Andrei Baroian

baroianandrei@yahoo.com | +40 765 469 415 | Leiden, Netherlands | Romanian

## About me

MSc Computer Science (AI) student at Leiden University, currently doing my thesis at SPY Lab (ETH Zurich) on prompt injection attacks against vision-language models. Past research includes LLM pre-training (architectural variants) and post-training (speeding up GRPO with Prompt Reuse), with smaller projects in quantization and mechanistic interpretability. Excited about autoresearch, scared of cybersecurity capabilities of agents.

## Projects & Research

**MSc Thesis: (Image) Prompt Injection** Feb 2026 – Present

SPY Lab, ETH Zurich — Supervised by Jie Zhang and Florian Tramèr. Studying transferability of image prompt injections: adversarial images optimized on open-source VLMs that transfer to closed-source models (GPT, Gemini, Claude). Concurrently exploring prompt injection attacks on OpenClaw agents, designing self-propagating worm attacks that modify agents' internal goal files and spread to other agents.

**Adversarial Attacks on VLA Models in Humanoid Robots** Jan 2026 – Present

ETH Robotics Club — Robotics Safety Division. Creating adversarial attacks against Vision-Language-Action (VLA) models in humanoid robots. Investigating how visual perturbations can override task instructions and induce harmful behaviors. Concurrently working on defenses.

**Prompt Replay: Speeding Up GRPO** arXiv:2603.21177

An overhead-free online data selection method for GRPO that reuses prompts (not trajectories) to preserve on-policy optimization. Buffers medium-difficulty prompts near a 50% pass rate to maximize learning signal. Tested on Llama-3.2-3B and Qwen3-8B, reducing zero-variance prompts and accelerating early training gains.

**Crown, Frame, Reverse: Layer-Wise Scaling Variants for LLM Pre-Training** arXiv:2509.06518

Explored architectural variants that redistribute capacity across transformer layers during pre-training. Introduced three layer-wise scaling patterns using linear interpolation of FFN widths and attention head counts. Pre-trained 180M parameter models on 5B tokens; all variants converged to better performance than an equal-cost isotropic baseline.

## Education

Aug 2024 – Jul 2026 **MSc Computer Science: Artificial Intelligence**, Leiden University, The Netherlands (GPA: 8.5/10)
Notable grades: Seminar in Deep Reinforcement Learning (10), Deep Learning (9.0), Seminar in Deep Learning (9.0), Natural Language Processing (9.0).

Feb 2026 – Jul 2026 **Exchange Semester — MSc Thesis at SPY Lab**, ETH Zurich, Switzerland
Adversarial attacks on vision-language models. Supervised by Jie Zhang and Florian Tramèr.

Sep 2021 – Jul 2024 **BSc Entrepreneurship & Business Innovation**, Tilburg University, The Netherlands

## Work Experience

**ETH Robotics Club, Zurich - Robotics Safety Researcher** Jan 2026 – Present

Part of the Robotics Safety division. Exploring adversarial attacks and defenses of VLA models in humanoid robots.

**Leiden University - Teaching Assistant, Automated Machine Learning** Sep 2025 – Jan 2026

– Grade assignments and provide feedback.

– Guide students in selecting, understanding, and presenting research papers.

**Akida, The Hague - Data Engineer** Mar 2025 – Present

– Develop filtering logic to detect construction projects in public-sector sources using heuristics & GenAI.

– Build the extraction pipeline for summarization and structured information retrieval with LLMs.

– Collaborate on annotation workflows and quality evaluation of LLMs.

## Skills

| | |
|---|---|
| LLMs | Pre-training (data prep, transformer design, distributed training), post-training (RLVR/GRPO, SFT), prompt engineering, quantization, mechanistic interpretability. |
| Adversarial ML | Prompt injection attacks (text & image), adversarial patches for VLMs/VLAs, transferability across models. |
| NLP | Named Entity Recognition, information retrieval, annotation workflows. |
| Frameworks & Tools | PyTorch; Hugging Face Transformers; Weights & Biases; OpenAI & Google Gemini APIs; Python; Linux; Slurm; Git; Docker; Azure. |