

Amir Nuriyev

+7-777-224-2032 | amir.ravnur@gmail.com | [LinkedIn](#) | [GitHub](#) | [Google Scholar](#) | anuriyev.com

EDUCATION

MBZUAI

B.S. in Artificial Intelligence

August 2025 – Present

Abu Dhabi, UAE

EXPERIENCE

Research Fellow

Heron AI Security

January 2026 – May 2026

Remote

- Lead a 5-person research team of cybersecurity professionals and PhDs
- Demonstrated distinguishability of expert activations from power traces

Research Scholar

ML Alignment & Theory Scholars

June 2025 – Present

Berkeley, CA

- AI Safety research program, working under the mentorship of RAND TASP Fellow Gabriel Kulp
- MATS Extension scholar, MoE adversarial attack research proposal ranked in top 15% of cohort

Research Fellow

Supervised Program for Alignment Research

February 2025 – May 2025

Remote

- Originated novel direction of RF and power side-channel attacks on MoE expert routing
- Finetuned MoE LLMs and built classifiers to emit and detect side-channel information

Data Science Intern

Kazakhtelecom

June 2023 – August 2023

Almaty, Kazakhstan

- Calculated the price reliefs for the victims of the Abay region wildfire, 600k+ affected
- Used ML models to find optimal cellular plan offering for 100k+ clients

SELECTED WORKS

A. Nuriyev, G. Kulp. Expert Selections In MoE Models Reveal (Almost) As Much As Text (2026)

ICLR Workshop on Trustworthy AI, SAGAI Workshop @ IEEE S&P (Oral)

- Demonstrated that MoE expert selections leak substantial information and can be decoded into text
- Trained a transformer-based decoder, reaching 91.2% top-1 accuracy on OpenWebText

ACHIEVEMENTS

Codeforces Master (top 3%)

IOI Team Selection Training (top 10 nationally)

International Zhautykov Olympiad (silver medal, top 70/300)

TALKS

Out-of-Band Verification Using Side Channels, Workshop on Assurance and Verification of AI Development (AViD)

TECHNICAL SKILLS

Languages: C/C++, Python, SQL

Developer Tools: Git, Docker, SLURM

Libraries: PyTorch, Transformers, Pandas, NumPy, Matplotlib