Andrii Nikolaiev, PhD

Cambridge, UK
(+44) 7768 773026
adnikolaiev@gmail.com
linkedin.com/in/andynik
Google Scholar
andynik.github.io

Summary

AI Researcher with a PhD in Computer Science and 5+ years of experience in AI/ML. Specialising in LLM benchmarking, human evaluations, and experimentation design. 10+ years in EdTech, leading educational programs and startup initiatives.

Education

Visiting Research Student
NLIP Group, University of Cambridge (2023–2025)

Ph.D. in Computer Science
Taras Shevchenko National Univ. of Kyiv (2020–2025)

B.Sc. & M.Sc. in Computer Science
Taras Shevchenko National Univ. of Kyiv (2014–2020)

Grants & Awards

- Emergent Ventures Fellow, Mercatus Center, USA (2023–2026)
- Fully funded research fellowship, University of Cambridge, UK (2023–2025)

Skills

- **LLM Evaluation & Research:** Benchmarking, CoT, LLM-as-a-Judge, Prompt Engineering, Human Evaluation, Rubric Design
- **Programming languages:** Python, SQL, R, C/C++, LaTeX
- **ML Frameworks & Inference:** PyTorch, Hugging Face, TensorFlow, vLLM, llama.cpp, Ollama, CUDA
- **Infrastructure:** HPC, Google Cloud Platform, Git, Linux/UNIX, Kubernetes, Docker
- **Natural languages:** English (fluent), Ukrainian (native)
- **Interests:** Half-marathons (PB: 01:31:16), football, hiking, yoga

Research & Professional Experience

Project Manager & Researcher, Mercor *Feb 2026 – Present*

- Developed evaluation metrics and designed batch-run experiments assessing frontier AI models on economically meaningful tasks across multiple professional domains for *APEX-Agents*, *Terminal-Bench* benchmarks.
- Analysed AI agentic behaviour – MCP tool usage, computer & browser user actions (CUA, BUA), reasoning, code execution.
- Oversaw large teams of domain experts (700+) across industries to design experimentation setups and rubrics aimed at revealing frontier AI models limitations.

Researcher in NLP, University of Cambridge *Apr 2023 – Present*

- Developed *Combi-Puzzles*, a novel symbolic benchmark with templated combinatorial tasks, revealing performance gaps (up to 30%) in reasoning accuracy across variations.
- Deployed and evaluated open-source and proprietary LLMs across HPC clusters and cloud-based inference environments.
- Designed and led human evaluation studies, comparing model responses against expert judgements.

Researcher in LLM Benchmarking, Osiris AI *Aug – Sep 2025*

- Built a multimodal OCR and segmentation pipeline for historical handwritten records, achieving a 15% accuracy gain via custom RAG methods and prompt engineering.

Ph.D. in Computer Science, Taras Shevchenko National Univ. *2020 – 2025*

- Thesis on LLM evaluation for abstract thinking and problem-solving, revealing differences in reasoning patterns between models and humans.
- Benchmarked 15+ language models on reasoning tasks over 860K+ Chain-of-Thought examples.
- Implemented LLM-as-a-Judge for data augmentation on textual data, reducing annotation costs while maintaining alignment with expert rubrics.

Selected Publications

- Neural Network Methods for Selecting and Generating Synthetic Variations of Combinatorial Problems, *Cybernetics and Systems Analysis*, Springer Nature, 2025.
- Can Language Models Rival Mathematics Students? Evaluating Mathematical Reasoning through Textual Manipulation and Human Experiments, *ACL RR*, arXiv preprint, 2024.

Leadership Experience

Co-founder & CEO, Kvanta.STEM EdTech Startup *2016 – Present*

- Led a national STEM year-round programme with 1,000+ learners and 40+ instructors across Olympiad-level mathematics, programming, and science.
- Secured $150K+ in funding; designed curriculum and educational annual boot camps for up to 250 participants (online & in-person).

Mentorship & Other Roles

- Managed national on-site educational programmes and $70K budgets at *NGO Kontora Pi*; coordinated institutional partnerships at *Roosh Investment Group*.
- Assisted *Machine Learning for Real-world Data* seminars at the *University of Cambridge*; *Databases* and *Cloud Computing* at *KNU*.