Grgur **Kovač**

AI Researcher

kovac.grgur@gmail.com | grgkovac.github.io | GitHub | LinkedIn | X (Twitter) | Bluesky | Google Scholar

AI researcher & engineer specializing in LLM robustness benchmarking, fine-tuning, autonomous exploration with 5+ years in research (≥200 citations) and 1-2 years in industry. An empirical position paper on LLM robustness/fuzzing (≥90 citations) extended to an LLM robustness leaderboard (≥33K visits). EMNLP 2025 Oral on data properties that mitigate degradation in iterative LLM finetuning/generation loops. Industry experience developing and improving production ML/DL models. Proficiency in vLLM, unsloth, PyTorch, hf datasets/transformers, cluster-scale job orchestration, experimental design and statistical analysis, with a proven ability to enter new domains, conceptualize, direct, and adapt projects on their frontier.

Skills and interests

Relevant Highlights

- **LLM benchmarking**: large scale LLM stability/fuzzing benchmarking; vLLM, HF transformers
- **LLM training and inference pipelines** data analysis for collapse in iterative LLM fine-tuning/generation loops (PEFT); vLLM, Unsloth, HF
- **Data Engineering**: scalable data cleaning, clustering, mixing, pipelines; HF_datasets, scikit-learn
- **Cluster Orchestration** for job workflows; SLURM on the Jean Zay supercomputer
- **AI agents:** Agentic RAG to synthesize/discuss my work; PydanticAI, and GCP;

Technical skills

- **Specializations**: LLM Robustness Evaluation & Benchmarking, LLM Fine-Tuning, Synthetic Data (Model Collapse), Training Data, Autonomous exploration
- **Programming**: Unsloth, HF transformers/datasets, PyTorch, sentence-transformers, scikit-learn, Python
- **Infrastructure**: vLLM, SLURM (large cluster), Docker, Git
- **Research**: Statistical Analysis (regression analysis, CFI, FDR, ANOVA), experimental design

Other

- **Languages**: English, Croatian, French (Conversational)
- **Hobbies:** Maintaining old/vintage bicycles

Key Achievements

Scientific Impact

- Influential paper on LLM robustness/fuzzing (**≥90 citations**)
- **EMNLP 2025 Oral** degradation in iterative LLM fine-tuning/generation loops
- Total **≥200 citations**

Practical Impact

- LLM robustness/fuzzing Leaderboard (**≥33K visits**)
- Training production models; e.g. text-based classifier (**≥200M requests/year**)

Experience

PhD Student | Flowers AI and CogSci Lab (INRIA)

Bordeaux (France) *Mar 2022 - Nov 2025*

- **LLM robustness benchmarking**: Created and maintained the StickToYourRole Leaderboard - evaluating LLM role-play value robustness to context change (fuzzing), validity and reliability analysis (**≥33K visits**, ∼1700/month); , ,
- **LLM robustness positioning**: presented and empirically supported a positioning of LLM context-sensitivity (**≥90 citations**); ,
- **Data analysis, Fine-tuning and & Model Collapse**: regression-based data analysis of properties related to degradation (quality, diversity, bias) of LLM generated content in iterative fine-tuning/generation loops (**EMNLP 2025, Oral**); ,
- **Socio-Cognitive AI**: designed a procedural interactive environment generator based to study generalization of socio-cognitive abilities in Deep RL and LLM-based agents; ,
- **Autonomous exploration (LLM & DRL)**: Creating an autonomous exploration architecture with LLM-based reward function generation for a DRL agent (co-author: conceptualization, brainstorming);
- **LLM bias drift**: uncovering attractors in iterative LLM-generated stories (**ICLR 2025**, second author: large-scale cluster experiment setup); ,
- design and initial development of a LLM-based qualitative analysis tool, validation/reliability with human annotations (taken on by another student);

Research engineer | Flowers AI and CogSci Lab (INRIA)

Bordeaux (France) *Nov 2019 - Jan 2022*

- **Autonomous exploration (VAE & Deep RL)**: augmenting novelty based goal driven exploration with Learning Progress for 3D vision-based DRL agents;
- **Socio-Cognitive AI**: creating interactive environments for Deep RL agents;

Research engineer (student internship) | Microblink

Zagreb (Croatia) *Jul 2017 - Sep 2019*

- **Computer Vision**: designed and trained production DL models for CV usecases (OCR, classification, ...)
- **NLP**: developed a text-based receipt classifier (classified **≥200M receipts/year**)
- **Efficiency**: creating a binary neural networks training method: large efficiency increase with negligible accuracy drops; NeurIPS 2019 MicroNet efficiency challenge **6th place** (from 19 participating teams)

Education

PhD in Computer Science

Flowers Team (INRIA) | University of Bordeaux *Mar 2023 - Oct 2025*

- **Thesis**: Building, evaluating and understanding socio-cultural AI: leveraging concepts and methods from human sciences;
- **Advisors**: Pierre-Yves Oudeyer (Flowers Team, INRIA), Peter Ford Dominey (CNRS)
- **Jury**: Maarten Sap (CMU), Jan Šnajder (UZG), Clémentine Fourrier (Hugging Face), Mehdi Khamassi (CNRS), Vered Schwartz (UBC)

Master of Computer Science

Faculty of Electrical Engineering and Computing | University of Zagreb *Oct 2017 - Jun 19*

- **Thesis**: Multiple Object Tracking based on Deep Learning;
- **Advisor**: Zoran Kalafatić
- **Relevant course project**: developed an NLP retrieval-based classifier;

Bachelor in Computer Science

Faculty of Electrical Engineering and Computing | University of Zagreb *Oct 2014 - Jun 17*

- **Thesis**: A framework for training feed-forward fully connected neural networks;
- **Advisor**: Zoran Kalafatić

Publications

Peer-reviewed

- Grgur Kovač*, Jérémy Perez*, Rémy Portelas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2025. Recursive Training Loops in LLMs: How training data properties modulate distribution shift in generated data? In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025, Oral)
- Jérémy Perez, Grgur Kovač, Corentin Léger, Cédric Colas, Gaia Molinaro, Maxime Derex, Pierre-Yves Oudeyer, and Clément Moulin-Frier (2025). 'When LLMs Play the Telephone Game: Cultural Attractors as Conceptual Tools to Evaluate LLMs in Multi-turn Settings'. In: *The Thirteenth International Conference on Learning Representations (ICLR 2025)*
- Grgur Kovač, Rémy Portelas, Masataka Sawayama, Peter Ford Dominey, and Pierre-Yves Oudeyer (2024b). 'Stick to your Role! Stability of Personal Values Expressed in Large Language Models'. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 46
- Grgur Kovač, Rémy Portelas, Masataka Sawayama, Peter Ford Dominey, and Pierre-Yves Oudeyer (Aug. 2024a). 'Stick to your role! Stability of personal values expressed in large language models'. In: *PLOS ONE* 19.8
- Grgur Kovač, Rémy Portelas, Peter Ford Dominey, and Pierre-Yves Oudeyer (2024). 'The SocialAI school: a framework leveraging developmental psychology toward artificial socio-cultural agents'. In: *Frontiers in Neurorobotics* Volume 18 - 2024
- Grgur Kovač, Adrien Laversanne-Finot, and Pierre-Yves Oudeyer (2022). 'Grimgep: learning progress for robust goal sampling in visual deep reinforcement learning'. In: *IEEE Transactions on Cognitive and Developmental Systems* 15.3

Preprints

- Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer (2023). 'Large language models as superpositions of cultural perspectives'. In: *arXiv preprint arXiv:2307.07870*

Workshops

- Grgur Kovač*, Rémy Portelas*, Katja Hofmann, and Pierre-Yves Oudeyer (June 2021). 'SocialAI 0.1: Towards a Benchmark to Stimulate Research on Socio-Cognitive Abilities in Deep Reinforcement Learning Agents'. In: NAACL. Accepted at NAACL ViGIL Workshop 2021. Mexico City, Mexico (Spotlight)
- Grgur Kovač, Rémy Portelas, Peter Ford Dominey, and Pierre-Yves Oudeyer (July 2023). 'The SocialAI School: Insights from Developmental Psychology Towards Artificial Socio-Cultural Agents'. In: TOM 2023 -First Workshop on Theory of Mind in Communicating Agents - ICML 2023 Workshop. Honolulu (Hawaii), United States
- Guillaume Pourcel, Thomas Carta, Grgur Kovač, and Pierre-Yves Oudeyer (2024). 'Autotelic LLM-based exploration for goal-conditioned RL'. In: Intrinsically Motivated Open-ended Learning Workshop at NeurIPS 2024

*equal contribution