

Greg (Grgur) Kovač

AI RESEARCHER

[✉ kovac.grgur@gmail.com](mailto:kovac.grgur@gmail.com) | [🏡 grgkovac.github.io](http://grgkovac.github.io) | [GitHub](https://github.com/grgkovac) | [LinkedIn](https://www.linkedin.com/in/grgkovac/) | [X \(Twitter\)](https://twitter.com/grgkovac) | [Bluesky](https://bluesky.social/@grgkovac) | [Google Scholar](https://scholar.google.com/citations?user=HdJyQAAJAAQ&hl=en)

AI researcher specializing in data and LLM Evaluation with 5+ years in research (190+ citations) and 1-2 years in industry. EMNLP 2025 Oral on identifying data properties that mitigate degradation in iterative synthetic data generation with LLMs, and an empirical position paper on LLM robustness (80+ citations). Created and maintained an LLM robustness and validity leaderboard (+30K visits). Proficiency in tools such as vLLM, SLURM, unsloth, hf transformers/datasets, PyTorch, and experimental design with statistical analysis. My experience has taught me how to quickly master and contribute to new domains.

Skills and interests

TECHNICAL SKILLS

- **Specializations:** Training Data, Synthetic Data (Model Collapse), LLM Robustness/Stability Evaluation & Benchmarking, Fine-Tuning
- **Programming:** Unsloth, HF transformers/datasets, PyTorch, Python
- **Infrastructure:** SLURM (large cluster), vLLM, 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

- EMNLP 2025 Oral on training data and synthetic data
- influential paper on LLM robustness (80+ citations).
- 190+ total citations

PRACTICAL IMPACT

- LLM Leaderboard attracting 31K+ 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 - Oct 2025

- **Synthetic Data & Model Collapse:** identifying data properties that foster or mitigate the degradation (quality, diversity, bias) of LLM generated data during iterative fine-tuning and synthetic data generation, i.e. model collapse (**EMNLP 2025, Oral**); [📄](#), [🔗](#)
- **LLM evaluation and benchmarking:** Created and maintained the StickToYourRole Leaderboard - evaluating LLM role-play value stability/robustness, validity and reliability of simulated populations (**31K+ visits**); [📄](#), [🔗](#), [🔗](#)
- **LLM evaluation and robustness:** presented and empirically supported a positioning of LLM context-sensitivity and culture expression (80+ citations); [📄](#), [🔗](#)
- **Socio-Cognitive AI:** designed a procedural interactive environment generator based on developmental theories to study generalization of socio-cognitive abilities in Deep RL and LLM-based agents; [📄](#), [🔗](#)
- uncovering attractors in iterative LLM-generated stories (**ICLR 2025**, second author: large-scale cluster experiment setup); [📄](#), [🔗](#)
- designing an LLM-based reward function generator (curriculum) for a DRL agent (co-author: conceptualization, brainstorming); [📄](#)
- 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

- **Deep RL & automatic curriculum learning:** creating an architecture augmenting novelty based exploration with Absolute Learning Progress for interactive vision-based Deep RL agents
- **Socio-Cognitive AI:** creating interactive environments for Deep RL agents; [📄](#)

Research engineer | Microblink / Photomath

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; [file](#)
- **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:** reimplementing and improving a Multiple Object Tracking approach; [file](#)
- **Advisor:** Zoran Kalafatić
- **Relevant course project:** developed an NLP retrieval-based classifier; [file](#)

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; [file](#)
- **Advisor:** Zoran Kalafatić

Publications

PEER-REVIEWED

- Grgur Kovač*, Jérémie 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)
- Jérémie 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