JAN FRANCISZEK PIOTROWSKI

janfpiotrowski (at) gmail (dot) com
jfpio.github.io ◇ github.com/jfpio

ABOUT ME

I am particularly interested in AI Safety, as I believe that artificial intelligence is a powerful tool but also a double-edged sword with the potential for serious risks if misaligned. Recently, I have been focusing on scalable oversight and mechanistic interpretability.

EDUCATION

University of Warsaw, Poland *2022-2024*
M.Sc. in Machine Learning

Thesis: Linking news to tweets cascades with contrastive learning approach
Overall GPA: 4.58 / 5

Sorbonne Université, France *2022 - Interrupted*
Postgraduate Studies in Project Management

Warsaw University of Technology, Poland *2018-2022*
B.Sc. in Computer Science

ACCEPTED PAPERS

Podolak, J., Łukasik, S., Balawender, P., Ossowski, J., **Piotrowski, J.**, Bąkowicz, K., Sankowski, P. (2024). "LLM-generated Responses to Mitigate the Impact of Hate Speech." EMNLP Findings 2024

EXPERIENCE

Wisent July 2025–August 2025
Machine Learning Research Engineer (Short-term) *Warsaw, PL*

- Built a reproducible LLM steering experimentation pipeline (HPO, evaluation harness, set up Wandb logging), revealing that syntax-level fixes did not translate into reasoning gains.
- Validated, fixed, and extended steering implementations; introduced CI and unit tests to stabilize rapid iteration.
- Released two steering-augmented models on Hugging Face; documented the evaluation protocol. Selected contributions: closed PRs by jfpio.

University of Warsaw January 2025 – June 2025
Machine Learning Researcher *Warsaw, PL*

- **Reinforcement Learning for Improving LLM Debate Protocols.**
Contributing to a research project that develops further the paper "Training Language Models to Win Debates with Self-Play Improves Judge Accuracy,". The project explores the prospects of RL training in multi-agent debate. I set up and managed LLaMA-family model runs on an HPC cluster via Singularity, and designed synthetic environments and proxy tasks to prototype debate dynamics. Collaborated on implementing novel reinforcement learning algorithms and new datasets.
- **Predicting Narcissism from User-Generated Texts (in collaboration with Faculty of Psychology).**
Initiated and led an interdisciplinary project to evaluate whether psychological traits-specifically narcissism-can be inferred from language patterns in user-generated texts. Developed a BERT-based regression model and a few-shot prompting setup using state-of-the-art LLMs. Compared model predictions against psychological benchmarks. Found that LLM-based few-shot methods lacked generalization, while transformer models showed potential but were limited by data scarcity. The project exposed core challenges in modeling personality traits and informed future approaches to data collection and model selection.

MIM Solutions January 2022 – December 2024
Machine Learning Engineer *Warsaw, PL*

- **Contrastive Learning for Linking Twitter Posts to Current Events (ERC-funded).** Led a research initiative, supervised by Prof. Sankowski, to develop contrastive learning methods using BERT-based dual encoders for linking Twitter posts with news articles in the context of real-world events. Personally handled experimental design, literature review, model development, and results analysis. Managed collaboration with University of Warsaw students. Findings were submitted to The Web Conference 2024; the revised manuscript is available as a preprint (arXiv:2312.07599).
- **Automated Counter-Speech to Mitigate Hate Speech (EMNLP Findings 2024).** Provided initial support for a student-led project employing LLMs to generate automated counter-speech against online hate speech targeting Ukrainian refugees. Supported project conceptualization, technical setup, and early experimentation. After two manuscript rejections, assumed lead role: introduced a novel engagement metric, expanded the literature review, added an ethical analysis, and rewrote the manuscript, leading to acceptance at EMNLP 2024 Findings (arXiv:2311.16905).
- **Photo Clustering Model for Major Advertising Firm (Project value: 25,000 EUR).** Directed development of an advanced photo clustering model utilizing embeddings, GroundingDINO, and GPT-4, achieving 99.5% accuracy (up from 95%). Oversaw client communication, requirements gathering, and project planning. Received positive public feedback from the client on LinkedIn.
- **Disinformation Analysis System for NASK.** Co-designed and implemented a machine learning system for detecting and analyzing online disinformation for NASK, Poland's national internet safety institute. Presented the solution at Warsaw Computer Science Days 2024, earning a 4.8/5 participant rating.
- **Public Writing on AI and Platform Governance.** Authored a public-facing article analyzing the societal risks of AI-driven content moderation, drawing from insights gained during work on disinformation and counter-speech systems. Published by Klub Jagielloński: Stanisław Lem przewidział patologie X-a.
- **Technical and Infrastructure Responsibilities.** Regularly trained, evaluated, and fine-tuned deep learning models on GPU clusters, using Hydra for configuration, Optuna for hyperparameter optimization, and Neptune for experiment tracking.

VOLUNTEERING

Polish Scouting Association (ZHR) January 2017 – Present

Scout Leader, Board Member – Mazovian Region

- Serve as Board Member for the Mazovian Region, responsible for oversight and quality assurance of summer and winter camps for over 2,000 youth participants annually.
- Directly managed camps of 120 participants (ages 11–20), including logistics, safety, and a 25,000 EUR budget; provided mentorship and support to team leaders.
- Led a local scout community for 4 years, expanding from 8 to 10 units and coordinating activities for 250 members; reduced adult leader turnover by 50% through new program development and mentorship of 30 leaders.
- Organized high-impact public discussions on youth and societal issues with leading Polish think tanks; the videos were published on (YouTube).