Nora Petrova AI Research Engineer

Last update: February 25, 2026 · Up-to-date version at https://nlpet.github.io/resume

CONTACT

London
github.com/nlpet
linkedin.com/in/nora-petrova
nora.axion@gmail.com

EDUCATION

BSc Mathematics & Physics (hons) — Open University, 2013–2018

BA Informatics — Molde University, 2007–2010

RESEARCH INTERESTS

AI Safety & Alignment

Mechanistic Interpretability

Adversarial Robustness & Red Teaming

Behavioural, HITL & Agentic Evals

Societal Impacts of AI

SKILLS

Python, PyTorch, Julia, JS/TS

React, Full Stack, AWS, GCP

LLMs, NLP, Model Training & Fine-tuning

PUBLICATIONS

Pressure Reveals Character — arXiv, 2026

Unpacking Human Preference for LLMs (HUMAINE) — ICLR 2026

Commercial Pressure Evals — IASEAI 2026

LAT Improves Representation of Refusal — ICLR 2025

Evaluating Synthetic Activations from SAE Latents — NeurIPS 2024

HACKATHONS

AI Manipulation Hackathon — Apart Research, 2026

Democracy x AI — Apart Lab, 2024

Testing LLMs for Autonomous Capabilities — Apart Lab & METR, 2024

Anthropic Claude Hackathon — Anthropic London, 2023

SUMMARY

AI Research Engineer building evaluation frameworks and interpretability tools to study how frontier models align with human values. My work spans agentic, human-in-the-loop, and behavioural evaluations of safety and alignment, mechanistic interpretability of model internals, and large-scale studies of how people experience and collaborate with AI systems. I am driven by the question of how we integrate AI into society responsibly, and I believe that rigorous evaluation, mechanistic understanding, and empirical study of human-AI interaction are the foundations for getting this right.

EXPERIENCE

Staff Research Engineer @ Prolific

May 2023 - Present

- Leading AI evaluation research, building agentic and human-in-the-loop evaluation frameworks for studying safety and alignment in frontier models
- Led HUMAINE, a large-scale evaluation of human experience in using AI — paper accepted at ICLR 2026
- Developed commercial pressure evaluations testing how frontier models respond when commercial objectives conflict with user safety, finding most models have no "red line" — won 1st place at the AI Manipulation Hackathon, presenting at IASEAI 2026
- Developed a behavioural alignment benchmark with an interactive leaderboard and co-authored Pressure Reveals Character
- Working on agentic evaluations for ethical decision making, mechanistic interpretability of refusal circuits and black-box prefix attacks in open source models, and studying how to decouple reasoning from knowledge using tiny recursive models (TRMs)
- Involved in cross-team projects including synthetic polling, agent detection, LLM usage detection, and multi-human-agent collaboration studies
- Earlier work included developing an open source social reasoning RLHF dataset and researching methods for aligning LLMs to human values

AI Safety Researcher @ LASR Labs

July 2024 - Sept 2024

Research into AI safety as part of the LASR Labs 12-week research programme. Investigated how synthetic activations composed of SAE latents compare to real model activations in GPT-2, measuring sensitivity via KL divergence of output logits. Found that while synthetic activations behave comparably to real ones under sparsity and geometric similarity metrics, real activations contain structural properties beyond independent components. Paper Evaluating Synthetic Activations composed of SAE Latents in GPT-2 accepted at NeurIPS 2024.

AI Safety Researcher @ Apart Research

May 2024 - Ongoing

Research fellowship (now ongoing collaboration) into AI safety alongside role at Prolific. Studied how latent adversarial training (LAT) affects how language models encode refusal. Found that LAT concentrates refusal representation into fewer SVD components, making models more robust against external attacks but paradoxically more vulnerable to self-generated attack vectors. Paper LAT Improves the Representation of Refusal accepted at ICLR 2025.

NLP Engineer / Julia Developer @ planting.space [contract]

May 2023 - Mar 2024

- Developed a knowledge representation system by leveraging insights from Bayesian Inference, Category Theory and Natural Language Processing, using the Julia programming language and Python for NLP
- Researched and implemented state of the art methods within NLP using LLMs, including advanced prompting techniques and fine-tuning of models
- Developed, deployed and maintained an internal search tool for finding relevant internal documents using semantic search on chunked vectorised documents. The tool integrated with multiple internal systems

Lead Software Engineer @ Signal.ai

Apr 2022 - May 2023

- Led a team of engineers and researchers developing ML models and services for mention detection, entity disambiguation, sentiment analysis, text classification, and zero/few-shot learning
- Researched state-of-the-art NLP techniques, prototyped applications, and developed production-ready pipeline services using transformer-based models
- Built a RAG prototype searching against a large Elasticsearch cluster of news documents to respond to arbitrary queries

Earlier Roles

Senior Product Engineer @ Primer.ai
2021 – 2022

ML Engineer / Researcher @ Prodo.ai
2017 – 2018

ML Engineer → Principal ML Engineer @ Datatonic
2018 – 2021

Full Stack Engineer @ R3PI, Digital Catapult, AXS
2012 – 2017