Nora Petrova

Machine Learning Engineer

Last update: September 13, 2024

Up-to-date version of CV is available at
https://nlpet.github.io/resume

Location	London
GitHub	github.com/nlpet
LinkedIn	linkedin.com/in/nora-petrova
Email	nora.axion@gmail.com

I am an experienced Software Engineer and Machine Learning Engineer specialising in NLP. Particularly interested in AI alignment problems and (mechanistic) interpretability research. I am motivated to bridge the gap between the capabilities of humans and AI agents, and am fully invested in working towards building a future in which machines and humans cooperate with and augment one another. I am driven by my curiosity to understand the inner workings of the brain and think that advances in AI will play an important role in providing valuable insights into this long-standing puzzle.

Projects

Democracy x AI Hackathon (May 2024)

Participated in a Hackathon organised by Apart Lab. I worked on a project that explored how using Sleeper Agents can have repercussions to democratic processes. More information about the project is available on the project page.

- NLP
- LLMs
- AI Safety

InterpVis Hackathon (Apr 2024)

Participated in a Hackathon organised by Apart Lab and LISA. The team I was in worked on visualising correlations between features across layers in an SAE trained on activations from the residual stream. We developed a prototype for visualising the relationships between features as a directed graph and linked features to Neuronpedia for further analysis. The code for the tool can be found in its GitHub repository.

- NLP
- LLMs
- Mechanistic Interpretability
- Visualisations

Testing LLMs for Autonomous Capabilities Hackathon (Mar 2024)

Participated in a Hackathon organised by Apart Lab and METR where the goal was to develop tasks to detect replication and R&D capabilities of LLMs. I worked on a task about capabilities of LLMs to introduce backdoors in datasets and to fine-tune other models on them. Based on research from the Sleeper Agents paper by Anthropic.

- NLP
- LLMs
- R&D

Anthropic Claude Hackathon (Nov 2023)

Participated in a Hackathon organised by Anthropic London. I developed an app that uses Claude to help us make sense of the News. Its aim is to introduce nuance and multiple perspectives, and enable us to entertain multiple points of view when we make up our minds about current events. GitHub Repository for the project.

- NLP
- LLMs
- Next.js
- React
- Javascript
- R&D

Professional Experience

July 2024 - Sept 2024

AI Researcher @ LASR Labs

Conducting research into AI safety as part of the LASR Labs 12 week research programme. The project involves studying activation space features and linking them to model behaviour. We're specifically investigating which activation directions are especially "sensitive" (as measured by KL divergence of the output logits) in LLMs, i.e. affect the model outputs more, and we want to test whether these directions are related to SAE features.

LLMs | AI Safety | Mechanistic Interpretability

May 2024 - Ongoing

AI Researcher @ Apart Lab

Conducting research into AI safety with Apart Lab part-time. Research area is activation steering and targeted / non-targeted latent adversarial training in relation to refusal of harmful requests.

NLP | LLMs | AI Safety

May 2023 - June 2024

AI Engineer / AI Consultant @ Prolific [contract]

- Advising Prolific on AI strategy and product development, primarily within NLP
- Researched, developed and open sourced a social reasoning dataset, for fine-tuning of LLMs using RLHF
- Researching state-of-the-art methods for aligning LLMs to human values, e.g. democratic-fine-tuning

NLP | LLMs | RLHF | Python | Typescript | Next.js | Vue | R&D | AWS | Product | Full Stack

May 2023 - Mar 2024

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

- 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

NLP | LLMs | Python | Julia | Javascript | React | AWS | Research | Full Stack

Apr 2022 - May 2023

Lead Software Engineer @ Signal.ai [full-time]

- Led a team of engineers and researchers with the aim of developing machine learning models and services for MD (mention detection), ED (entity disambiguation), detecting sentiment, text classification, zero/ few shot learning approaches and custom models relevant to our clients
- Involved in researching relevant state-of-the-art techniques within NLP, prototyping applications, developing production ready pipeline services and optimising models for production
- Involved in bringing state-of-the-art approaches within NLP to client projects (transformer based models)
- Responsibilities included maintenance of existing pipeline services, roadmap for the team, strategy from a product point of view and continuous improvement of services
- Built a RAG prototype which searched against a large Elasticsearch cluster of news documents with the goal to respond to arbitrary queries

NLP | Python | Clojure | R&D | Modelling | Terraform | MLOps | AWS | Full-Stack | Management | Tech Lead

May 2021 - Mar 2022

Senior Product Engineer @ Primer.ai [full-time]

- Involved in building NLP powered applications in a full stack capacity (React, Typescript, Python)
- Responsible for performing data analysis, modelling and preparation, and coming up with machine learning approaches (within NLP) to client problems
- Involved in bringing state-of-the-art approaches within NLP to client projects (transformer based models)
- Researched appropriate storage layers for knowledge graphs, as well as techniques for building them (Neo4J, MongoDB, Postgres)

NLP Python Full Stack R&D Product

Nov 2019 - May 2021

Principal Machine Learning Engineer @ Datatonic [full-time]

- Led multiple client projects from a technical standpoint; providing direction for the technical delivery and the architecture of machine learning based solutions on Google Cloud Platform
- Performed research and applied state-of-the-art machine learning models and techniques in client projects
- Mentored more junior members of the team and managed a small team
- Participated in workshops and presented at client facing events
- Completed Google's Professional Machine Learning Engineer and Cloud Architect certifications

ML Python Full Stack R&D GCP MLOps Modelling Productionisation Management Client Projects

Sept 2018 - Nov 2019

Senior Machine Learning Engineer @ Datatonic [full-time]

- Worked on client projects to build and deploy machine learning models on Google Cloud Platform. This usually involved performing data analysis, building data processing pipelines, researching and developing machine learning models, and delivering solutions in short timeframes
- Worked on a variety of proof-of-concept projects for clients in problem areas such as NLP, document classification, computer vision, anomaly detection, recommender systems, and forecasting. The technical stack consisted of Python, Tensorflow / Keras / scikit-learn on GCP
- Led an internal project for standardising machine learning approaches and defining best practices
- Completed Google's Professional Data Engineer and Cloud Architect certifications

ML Python Javascript React Full Stack R&D GCP MLOps Modelling Productionisation Client Projects

Jun 2017 - Sept 2018

Machine Learning Engineer / Researcher & Full Stack Software Engineer @ Prodo.ai [full-time]

- Worked on developing deep learning models by leveraging the latest research in AI within graph based neural networks. Primarily using Python, PyTorch and NumPy for model building; pandas, scikit-learn, matplotlib for data, evaluation and visualisation.
- Worked across the stack on building an application which communicates machine learning insights to clients using Javascript, HTML/CSS, NodeJS, Koa, React.js, Redux, Webpack, Ava, RabbitMQ, Kubernetes, MySQL, Elasticsearch, docker, AWS, Google Cloud and others.
- Developed infrastructure tools for running machine learning experiments using primarily Python and AWS.

ML Python Typescript React Full Stack R&D AWS

Feb 2017 - Jun 2017

Full Stack Software Engineer @ R3PI [full-time]

- Built a suite of micro services which provide metrics and analytics, enabling clients to make better informed decisions about the vehicles under their control. The technologies we used were: Javascript, NodeJS, MySQL, MongoDB, Docker, Mocha, Chai, Istanbul, Bluebird, Hapi, Joi, Bunyan, Bookshelf, Knex, Mongoose among others.
- Project was completed in May and was deployed and used by a major client

Javascript NodeJS React Docker MongoDB TDD

Nov 2015 - July 2016

Full Stack Software Engineer @ Digital Catapult [full-time]

- Delivered an open source project which is an implementation of an open online digital rights service. The back-end was written in Python using the Tornado async framework. We also used Django, Flask, behave, pytest, Blazegraph (graph db using RDF/SPARQL), MySQL and others. The front-end was written in Javascript with React.js, Bacon.js, gulp and others.
- Worked on designing, building, testing and deploying the above mentioned project and helped on building an openstack cluster for a project that enables data scientists to mix closed data sets in a secure environment. Technologies used: Openstack, Hadoop, Ambari, Terraform, vagrant, docker, chef and AWS.

Javascript NodeJS React Python AWS Terraform Hadoop MySQL SPARQL

Dec 2012 - Nov 2015

Full Stack Software Engineer @ AXS.com [full-time]

- Worked on building a ticketing application which was to replace the existing system and make it easier to maintain, extend, integrate analytics and modernise the existing experience for our customers. The front-end was built using Javascript and React.js, and the back-end was written in Python. I worked across the stack.
- Implemented core functionality and helped with the roll-out of the product

Javascript React Python MySQL Full Stack

Specialisations & Courses

AI Safety Fundamentals Course - **BlueDot Impact** - 2024
NLP Specialisation - **Coursera** (deeplearning.ai) - 2020
Professional ML Engineer - **Google Cloud** - 2020
Deep Reinforcement Learning Engineer Specialisation - **Udacity** - 2018
Machine Learning Engineer Specialisation - **Udacity** - 2018

Education

BSc Mathematics & Physics (hons) - **Open University** - 2013 - 2018
BA Informatics - **Molde University** - 2007 - 2010