Friederike VP-of-AI-Hofmair

Senior AI Research Scientist

friederike.hofmair@example.de
Saarbruecken, Deutschland
linkedin.com/in/friederike-hofmair
+49 681 4485 2270
friederike-hofmair.science
scholar.google.com/citations?user=friederike-hofmair

PROFIL

AI Research Scientist mit Promotion Dr. rer. nat. (MPI Tuebingen, summa cum laude) und 6 Jahren Forschungs-Erfahrung am DFKI Saarbruecken und Helmholtz AI Munich. Schwerpunkt Foundation Models, RLHF und Mechanistic Interpretability mit 18 Publikationen bei NeurIPS, ICML, ICLR und ACL. Heisenberg-Stipendium-Empfaengerin der DFG und ELLIS Member.

BERUFSERFAHRUNG

Senior AI Research Scientist 07/2022 - heute

DFKI Deutsches Forschungszentrum fuer Kuenstliche Intelligenz

Saarbruecken + Berlin, Deutschland

Senior-Forscherin im Smart-Data-und-Knowledge-Services-Lab

- Leitung einer 4-koepfigen Forschungsgruppe zu Mechanistic Interpretability fuer deutschsprachige LLMs
- Einwerbung von 480.000 EUR DFG-Drittmitteln und 1,2 Mio. EUR BMBF-Verbund-Foerderung in 18 Monaten
- Publikation von 8 Papers bei NeurIPS, ICML und ACL als Erst- oder Letzt-Autorin (h-Index 18)
- Co-Organisation des NeurIPS 2024 Workshops on Interpretable Foundation Models in Vancouver

AI Research Scientist (Postdoc) 10/2020 - 06/2022

Helmholtz AI Munich Muenchen, Deutschland

Postdoctoral Researcher im Helmholtz-AI-Konsortium an RLHF-Methoden

- Forschung zu Reward-Modelling und RLHF-Stabilitaet auf 480.000 menschlichen Praeferenz-Paaren
- Co-Autorin von 6 Papers bei NeurIPS, ICLR und ICML in 22 Monaten
- Mentoring von 3 Doktoranden und 4 Masterstudenten an der TUM und LMU
- Praesentation auf Konferenzen in Vancouver, Wien und Singapur als Invited Speaker

AUSBILDUNG

Dr. rer. nat. Maschinelles Lernen 10/2016 - 09/2020

Max-Planck-Institut fuer Intelligente Systeme + Universitaet Tuebingen

Tuebingen + Stuttgart, Deutschland

Maschinelles Lernen summa cum laude

M.Sc. Computer Science 10/2013 - 09/2016

ETH Zuerich Zuerich, Schweiz

Computer Science (Schwerpunkt ML) GPA: 1,0

FÄHIGKEITEN

- Python (PyTorch, JAX, Triton CUDA)
- Probabilistische ML & Bayesian Methods
- Reinforcement Learning + RLHF
- Foundation Models + Mechanistic Interpretability
- Wissenschaftliches Schreiben (LaTeX, Overleaf)
- Hugging Face Stack
- Slurm + DGX SuperPOD
- Citation-Tracking + Scholar Profile

PROJEKTE

interp-llm-de (DFG-Projekt) — 01/2024 - heute

Mechanistic Interpretability fuer deutschsprachige LLMs, 480.000 EUR Drittmittel, 4 Doktoranden, 8 Papers in Vorbereitung

RLHF for Industrial Robotics (BMBF) — 06/2022 - 12/2023

RLHF-Pipeline fuer Industrieroboter bei Bosch, Erfolgsrate 0,68 auf 0,89, 3 Papers bei NeurIPS

ZERTIFIKATE

- Heisenberg-Stipendium DFG — 06/2024
- ELLIS European Lab for Learning + Intelligent Systems Member — 11/2023
- BMBF Forschungsverbund AI Center Tuebingen-Stuttgart-Muenchen — 03/2022
- NeurIPS / ICML / ICLR / ACL Co-Autorin (18 Publikationen) — 11/2020

SPRACHEN

Deutsch	Muttersprache
Englisch	C2
Franzoesisch	B2

PUBLIKATIONEN

- Mechanistic Interpretability of German Language Foundation Models — 12/2024
- Stability and Convergence Properties of Reward Models in RLHF — 07/2024
- Reproducible Benchmarks for German Language Understanding — 05/2024

STÄRKEN

Forschungs-Sorgfalt

Bestehe auf reproduzierbaren Experimenten mit fixierten Seeds, Power-Analysen und Pre-Registration vor Papers

Tiefe in Foundations

Verstehe sowohl die Theorie (Information Geometry, Optimal Transport) als auch Engineering-Realitaet

Mentoring-Wirkung

Habe 6 Doktoranden betreut, 4 davon haben in Top-Konferenzen publiziert und sind in Industrie-Forschung gewechselt