

JEREMY HERBST

Website ◊ Google Scholar ◊ GitHub ◊ LinkedIn

Phone: (+49) 1602718834 ◊ Email: jeremy.herbst111@gmail.com

EDUCATION

University of Copenhagen (UCPH) PhD student Mechanistic Interpretability for LLM Security	09/2026 – Current
University of Hamburg (UHH) M.S. in Informatics	04/2023 – 06/2025
University of Hamburg (UHH) B.S. in Software Systems Development	10/2019 – 03/2023

RESEARCH INTERESTS

Interpretability, Explainable AI, AI Security/Safety, Modular AI

RESEARCH

Interpreting Mixture-of-Experts Language Models at Expert Level Supervisors: Dr. Jae Hee Lee , Prof. Dr. Stefan Wermter	06/2025 – 09/2026 UHH
– Studied interpretability of MoE experts in LLMs [1] – Discovered that MoE experts are less <i>polysemantic</i> than dense feed-forward networks by using k-sparse probing – Provided new perspective on model interpretability and expert specialization using automatic interpretability and causal attribution – Investigated alternative interpretability techniques by training Sparse Autoencoders (SAEs)	
Master’s Thesis – Expert Specialization in Mixture-of-Experts Models Supervisors: Dr. Jae Hee Lee , Prof. Dr. Stefan Wermter	11/2024 – 06/2025 UHH
– Analyzed expert specialization in OLMoE (7B) across depth – Used activation statistics, clustering, similarity metrics to identify specialization patterns – Performed causal router interventions to verify expert-specific behavior	

PUBLICATIONS

- [1] Jeremy Herbst, Stefan Wermter, and Jae Hee Lee. “The Expert Strikes Back: Interpreting Mixture-of-Experts Language Models at Expert Level”. In: *Forty-third International Conference on Machine Learning*. 2026. URL: <https://openreview.net/forum?id=npM0aMWWrW>.

REVIEWING

Mechanistic Interpretability Workshop	ICML 2026
Actionable Interpretability Workshop	COLM 2026

WORK EXPERIENCE

crossnative Working student – Data Engineering Built data pipelines using Snowflake to extract data for machine learning prototypes	04/2023 – 06/2024
PPI AG Working student – Web Applications Maintained web applications in a small team to deliver software updates	12/2021 – 04/2023
PPI AG Internship – Software Development	06/2021 – 10/2021

SKILLS

Languages	German: <i>Native</i> English: <i>Fluent</i> Spanish: <i>Intermediate (B1)</i>
Programming Languages	Python, Java, Rust, TypeScript/JavaScript, HTML
Tools	PyTorch, NumPy, scikit-learn, Hugging Face, Weights & Biases, CUDA

INTERESTS

Resistance training, handball, hiking

ACADEMIC REFEREES

Prof. Dr. Stefan Wermter

Professor of Computer Science, Chair for Knowledge Technology
University of Hamburg
Email: stefan.wermter@uni-hamburg.de

Dr. Jae Lee

Principal Investigator
University of Hamburg
Email: jae.hee.lee@uni-hamburg.de

Dr. Tom Petersen

Research Associate
University of Hamburg
Email: tompetersen@mailbox.org