Jakob Paul Zimmermann

jakob.paul.zimmermann@gmail.com

Website · Google Scholar · LinkedIn

Profile

M.Sc. Computer Science student at TU Berlin with bachelor's degrees in Mathematics and Computer Science; first author of three ML papers, including a CVPR 2026 Highlight. I develop methods for monitoring and analysing models through their internal representations and computations. My work spans representation-based runtime monitoring, game-theoretic transformer attribution, reinforcement-learning fine-tuning of language models, and machine-checked analysis of neural-network loss landscapes.

Research

Knowledge-Guided Failure Prediction — runtime monitoring of safety-critical misses **Fraunhofer IOSB (ML4Safety)**
SAIAD Workshop at CVPR 2026 · *first author* · *paper* *Fall 2024*

- Formulated the failure mode — a detector silently missing a safety-critical object — and built an end-to-end runtime monitor comparing a detector's internal features against foundation-model (DINO) embeddings, with no ground truth at inference time. Developed the approach, wrote the code, ran the experiments and drafted the paper.
- Raised person recall from **64.3% to 84.5% at a fixed 5% false-positive rate** on COCO, reported as an operating-point trade-off rather than an aggregate score.

Playing the Network Backward: A Game Theoretic Attribution Framework **Fraunhofer HHI**
Preprint, arXiv:2605.06212 · *first author, with J. Berend, G. Loho, S. Lapuschkin and W. Samek* *Feb 2026 – present*

- Originated and designed a framework that represents the backward pass as a two-player game over an extended computational graph, inducing *distributions over computation trajectories* and recovering gradient, LRP, and transformer attribution rules as special cases of a common formalism.
- Implemented and evaluated the transformer adaptation on **ViT-B/16**, improving over prior transformer-specific attribution methods.
- *Ongoing:* testing whether trajectory distributions identify computational pathways and component interactions with causal or intervention-relevant meaning.

Hidden Monotonicity: Explaining Deep Neural Networks via their DC Decomposition **FU Berlin**
CVPR 2026 (Highlight) · *first author, with G. Loho* · *paper* · *code* *Jul 2025 – present*

- Proved that any ReLU network decomposes into a difference of two *monotone convex* Maxout networks, using the polyhedral geometry (Newton polytopes) of network functions, and turned that structure into two implemented attribution methods, SplitCAM and SplitLRP.
- Benchmarked in PyTorch across VGG16 and ResNet18 on ImageNet-S using the Quantus evaluation suite, outperforming widely used attribution baselines on faithfulness, localisation and robustness metrics — e.g. pointing-game score **0.938 vs. 0.887** for the strongest baseline. Paper selected as a CVPR Highlight.

Reinforcement-Learning Fine-Tuning of Language Models **Fraunhofer HHI**
Work in preparation *2026*

- Designing and running fine-tuning experiments on the reward signals used for reinforcement learning on open-ended tasks, building the training and evaluation pipeline and benchmarking against standard policy-optimisation baselines across several model scales. In preparation.

Formal Analysis of Neural-Network Loss Landscapes
Lean 4 / mathlib · *machine-checked* · *with M. Grillo* · *in preparation* *2026 – present*

- Formalising the optimisation landscape of shallow ReLU networks in Lean 4, with machine-checked structural results on the critical points of the population loss. Investigating what they imply for optimisation dynamics and learned representations. Paper in preparation.

Activation Steering in Language Models **Independent**
Independent exploration *2026*

- Locating behaviourally relevant directions in a model's activations and intervening along them, to test how robustly the corresponding behaviour is encoded internally rather than expressed only at the output. Ongoing.

Education

Technische Universität Berlin — M.Sc. Computer Science **Berlin** *Oct 2024 – Mar 2027 (expected)*
Master's thesis completed; remaining requirements are coursework only.

Karlsruhe Institute of Technology (KIT) — B.Sc. Computer Science and **B.Sc. Mathematics** **Karlsruhe**
Coursework: Statistical Learning, Convex Geometry, Random Graphs, Graph Theory, Combinatorics *2020 – Mar 2024*

- Thesis *Induced Turán problems* → *Induced Turán problem in bipartite graphs* (with M. Axenovich), *Discrete Applied Math.* 360:497–505, 2025; follow-up *Bipartite Turán problem on cographs* (sole author, preprint, 2026).

Technical Skills

Core: Python (PyTorch, NumPy, scikit-learn, pandas), Lean 4, Git, Linux, LaTeX; also C++, Java, SQL.

Models & training: LLMs (RL fine-tuning), ViTs, DINO, CLIP, YOLO; probing and internal-representation analysis, monitor design and FPR-controlled evaluation.

Infrastructure: GPU-cluster experimentation — model and data sharding, data-loading throughput optimisation, multi-run experiment management. **Languages:** German (native), English (fluent).

Additional

Talks & presentations

- *From DC Decompositions to Two-Player Games: A Structural Approach to Explainability in Deep Learning* — invited talk, Hanno Gottschalk's research seminar (IMOS), TU Berlin — Jul 2026
- *Hidden Monotonicity* — Highlight poster, CVPR 2026, Denver — Jun 2026
- *Knowledge-Guided Failure Prediction* — talk, SAIAD Workshop at CVPR 2026, Denver — Jun 2026
- *Hidden Monotonicity* — poster, Workshop on Polyhedral Geometry for Neural Networks, Nuremberg — Mar 2026
- *Bipartite Turán problem on cographs* — talk, Szabó research seminar, FU Berlin — Jan 2026

Teaching: Tutor & Grader, Graph Theory (Fall 2023) and Stochastics (Spring 2023), KIT Karlsruhe.