

Jakob Paul Zimmermann

✉ jakob.paul.zimmermann@gmail.com 🌐 jakob-paul-zimmermann.com

Permitted to work full-time while completing my M.Sc.; willing to relocate.

Last updated: August 2026

Research Profile

I develop methods for monitoring and analysing models through their internal representations and computations: polyhedral geometry of network functions, difference-of-convex decompositions, game-theoretic transformer attribution, representation-based runtime monitoring, RL fine-tuning of language models, and machine-checked analysis of loss landscapes. First author of three ML papers, including a CVPR 2026 Highlight.

Research Experience

Research Assistant, Fraunhofer HHI

Berlin, February 2026 – present

- *Transformer interpretability with professor Wojciech Samek. Originated “Playing the Network Backward”, a game-theoretic attribution framework representing the backward pass as a two-player game on an extended computational graph. Also running RL fine-tuning experiments on language models (in preparation).*

Independent research

2026 – present

- *Formal analysis of neural-network loss landscapes in Lean 4 (with Moritz Grillo): machine-checked structural results on the critical points of the population loss of shallow ReLU networks. Paper in preparation. Also independent activation-steering experiments in language models.*

Research Assistant, Discrete Geometry Group (Free University Berlin)

Berlin, July 2025 – present

- *With professor Georg Loho, investigating the geometry of the Newton polytopes of neural networks and Difference-of-Convex decompositions. This structure underpins our CVPR 2026 Highlight “Hidden Monotonicity”.*

Research Assistant / Intern, ML4Safety (Fraunhofer IOSB)

Karlsruhe, Fall 2024

- *Runtime monitoring of image-recognition models and representation learning. Developed a knowledge-guided monitoring approach operating on internal representations and foundation-model embeddings such as DINO.*

Education

Technical University Berlin

Berlin

- M.Sc. in **Computer Science**
- Master’s thesis completed; remaining requirements are coursework only.

October 2024 – March 2027 (expected)

Karlsruhe Institute of Technology (KIT)

Karlsruhe

- B.Sc. in **Computer Science**
- B.Sc. in **Mathematics**
 - Coursework: Statistical Learning, Convex Geometry, Random Graphs and Networks, Graph Theory, Combinatorics.
- Bachelor thesis “Induced Turán problems” under supervision of Maria Axenovich and Thorsten Ueckerdt, resulting in the published paper “Induced Turán problem in bipartite graphs” (Discrete Applied Mathematics, 2025).

October 2020 – March 2024

April 2021 – March 2024

Haus Lindenhof Foundation

Schwäbisch Gmünd

Voluntary social year in a residential group for people with disabilities

March 2020 – August 2020

Teaching

Tutor & Grader, Graph Theory (KIT Karlsruhe)

Fall 2023

Tutor & Grader, Stochastics (KIT Karlsruhe)

Spring 2023

Skills

- **Programming:** Python (ML: PyTorch, NumPy, pandas, scikit-learn), C++, Java, $\text{\LaTeX}$, SQL, Lean 4 (formal verification). **Familiar with:** Kotlin, R, MATLAB, Haskell, Lisp.
- **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.
- **Other:** Git, Linux, parallel computing. **Languages:** German (native), English (fluent).