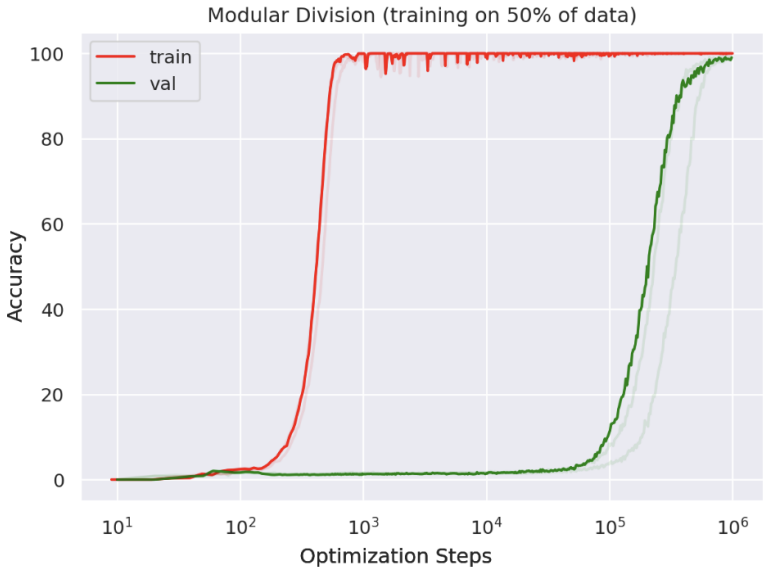
TECHNISCHE UNIVERSITÄT DARMSTADT

Thesis (B.Sc. / M.Sc.)
Investigating Grokking Phenomena: Neural Network Learning Dynamics through Sparse Crosscoders

Background: Grokking is a fascinating phenomenon in deep learning where neural networks initially memorize training data before suddenly generalizing to unseen examples after extended training [1]. This behavior, characterized by a significant gap between training and validation performance that unexpectedly closes after many optimization steps, challenges our understanding of generalization in neural networks. This thesis aims to investigate the underlying mechanisms of grokking using interpretable neural network architectures such as Sparse Crosscoders [2]. By leveraging these recently developed interpretable models, we can gain insights into how neural networks transition from memorization to genuine understanding of the underlying patterns, particularly in algorithmic tasks such as modular arithmetic.



Learning curve for modular division showing the grokking phenomenon. The model quickly achieves near-perfect training accuracy (red) while validation accuracy (green) remains low until suddenly improving after approximately 10^5 optimization steps. Image adapted from Power et al. [1].

Department 18
Electrical Engineering and Information Technology
Self-Organizing Systems Lab

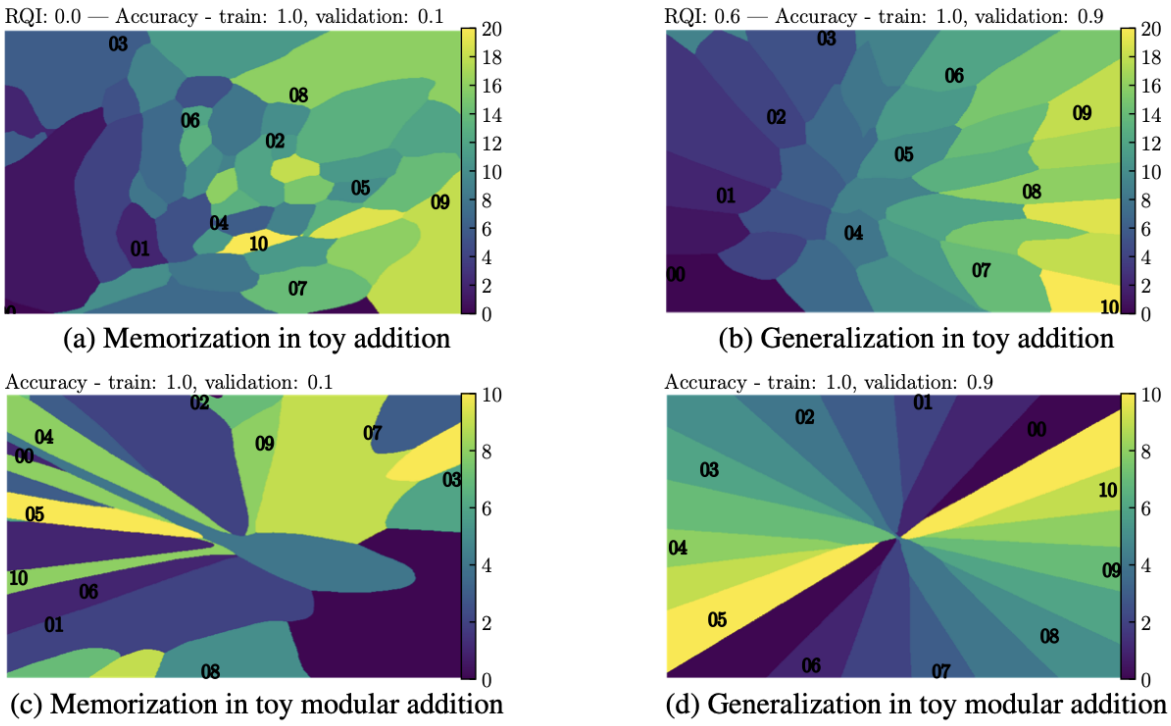
Prof. Dr. Heinz Koeppl
Head of lab

Philipp Froehlich
Project supervisor

S3|06 206
Merckstrasse 25
64283 Darmstadt

philipp.froehlich@tu-darmstadt.de
https://www.bcs.tu-darmstadt.de

December 25, 2025



(a) Memorization in toy addition (b) Generalization in toy addition

(c) Memorization in toy modular addition (d) Generalization in toy modular addition

Image from [4]. Visualization of the learned set of embeddings (p= 11). Grokking corresponds to learning meaningful embeddings.

Objective: Join our interdisciplinary team and advance the understanding of neural network learning dynamics through the lens of interpretability. We offer exciting thesis opportunities in the following areas:

- **Grokking Analysis:** Investigate the conditions under which grokking occurs across different tasks and model architectures, with a focus on mathematical and algorithmic problems.
- **Interpretability Methods:** Apply and extend interpretable neural network architectures like Sparse Crosscoders [2] to visualize and understand the internal representations before, during, and after the grokking transition.
- **Theoretical Modeling:** Develop theoretical frameworks to explain the sudden phase transition from memorization to generalization, potentially drawing connections to concepts from statistical physics and building upon work on implicit regularization in gradient descent [3].
- **Curriculum Design:** Explore how different training curricula and optimization strategies can accelerate or control the grokking process for more efficient learning.

Prerequisites:

- Background in computer science, mathematics, physics, electrical engineering, or related fields.
- Strong programming skills in Python and experience with deep learning frameworks (PyTorch preferred).
- Interest in neural network interpretability and machine learning theory.

- Basic understanding of information theory and statistical learning concepts is beneficial but not mandatory.

For further information, please contact Philipp Froehlich.

References

[1] Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). *Grokking: Generalization beyond overfitting on small algorithmic datasets*. arXiv preprint arXiv:2201.02177.

[2] Lindsey, J., Templeton, A., Marcus, J., Conerly, T., Batson, J., & Olah, C. (2024). *Sparse Crosscoders for Cross-Layer Features and Model Diffing*. Anthropic. https://transformer-circuits.pub/2024/crosscoders/index.html

[3] Soudry, D., Hoffer, E., Nacson, M.S., Gunasekar, S., & Srebro, N. (2018). *The Implicit Bias of Gradient Descent on Separable Data*. Journal of Machine Learning Research, 19(1), 1-57. https://www.jmlr.org/papers/volume19/18-188/18-188.pdf

[4] Liu, Z., Kitouni, O., Nolte, N. S., Michaud, E., Tegmark, M., & Williams, M. (2022). *Towards understanding grokking: An effective theory of representation learning*. Advances in Neural Information Processing Systems, 35