Can Demircan

Phd student interested in LLM interpretability, AI alignment, and representation learning.

can.demircan@tum.de Scholar candemircan OSF
@can_demircann @candemircan.bsky.social

Education

2023 – … **Ph.D., Technical University of Munich** in Machine Learning & Cognitive Science.
Supervisors: Dr. Eric Schulz (Helmholtz Munich) & Prof. Dr. Zeynep Akata (TUM).
Topics: AI alignment, LLM interpretability, and representation learning.

2021 – 2023 **M.Sc., University of Tuebingen** in Neural & Behavioural Sciences (Final Grade: 1.3).
Thesis: *Using tools of neuroscience to understand large language models*.

2017 – 2020 **B.A., Wadham College, University of Oxford** in Experimental Psychology.
Graduated with 1^{st} Class Honours

Experience

2022 – … **Instructor with the Software Carpentry**.
Taught and helped with workshops on Python, Git, and Bash at the University of Tuebingen and the University of Twente.

2021 – 2023 **Research Assistant** at the Max Planck Institute for Biological Cybernetics.
Investigated the representational basis of human learning in naturalistic tasks using online experiments and computational modelling.

Skills

Technical Python, R, HTML/CSS/JS/jsPsych, Git, Bash, Docker/Singularity, CI/CD, LaTeX.

Research Multi-device model training in PyTorch, online behavioural experiments, cognitive and statistical modeling, verbal, written, & visual communication of research.

Languages Turkish (native), English (fluent), German (beginner-intermediate).

Awards & Grants

2025 **Lambda Research Grant** $1000 worth of cloud compute for interpretability research on language models trained to capture human cognition.

2021 – 2023 **International Max Planck Research School Stipend** Two years of full funding to pursue an M.Sc. degree in Tuebingen.

Peer-Reviewed Publications & Preprints

2025 Saanum, T.*, **Demircan, C.***, Gershman, S. J., & Schulz, E. A circuit for predicting hierarchical structure in-context in Large Language Models. *Under Review.* [paper] [code]

Binz, M., …, **Demircan, C.**, …, Schulz, E. A foundation model to predict and capture human cognition. *Nature*. [paper]

Demircan, C.*, Saanum, T.*, Jagadish, A. K., Binz, M., & Schulz, E. Sparse Autoencoders Reveal Temporal Difference Learning in Large Language Models. *International Conference on Learning Representations (ICLR).* [paper]

2024 **Demircan, C.**, Saanum, T., Pettini, L., Binz, M., Baczkowski, B. M., Doeller, C., Garvert, M. M., & Schulz, E. Evaluating alignment between humans and neural network representations in image-based learning tasks. *Advances in Neural Information Processing Systems (NeurIPS).* [paper] [code] [data]

Özdemir, Ş., Şentürk, Y. D., Ünver, N., **Demircan, C.**, Olivers, C. N. L., Egner, T., & Günseli, E. Effects of context changes on memory reactivation. *Journal of Neuroscience.* [paper]

Şentürk, Y. D., Ünver, N., **Demircan, C.**, Egner, T., & Günseli, E. The reactivation of task rules triggers the reactivation of task-relevant items. *Cortex*. [paper]

Non-Archival Publications

2025 Naranjo, I., **Demircan, C.** , Schulz, E. How Does an LLM Process Conflicting Information In-Context? 8^{th} *Annual Conference on Cognitive Computational Neuroscience (CCN).* [paper]

2022 **Demircan, C.**, Pettini, L., Saanum, T., Binz, M., Baczkowski, B. M., Doeller, C., Garvert, M. M., Schulz, E. Decision-Making with Naturalistic Options. *In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 44, No. 44)* [paper] [code & data]

Supervision

2024 – 2025 **Ivan Naranjo** B.Sc. in Computer Science, Technical University of Munich.
Thesis: *How Does an LLM Process Conflicting Information In-Context?*

Reviewing

2025
- International Conference on Learning Representations (ICLR)
- Conference on Neural Information Processing Systems (NeurIPS)
- NeurIPS Workshop on CogInterp: Interpreting Cognition in Deep Learning Models
- Conference on Cognitive Computational Neuroscience (CCN)