

Mats Leon Richter

Research Scientist | Pretraining, Posttraining and Agentic AI

Berlin, Germany | +49 151 291 036 27 | matsleonrichter@gmail.com | [LinkedIn](#) | [Google Scholar](#)

Research scientist focused on effective, efficient training at scale: architectures, pretraining, posttraining and synthetic data. Equal-contribution first author and oral presenter of Würstchen at ICLR 2024. Contributed to Holo and Holotron at H Company; previously Mila and ServiceNow Research.

Large-scale training: Trained 395B-parameter MoE and 72B-parameter dense attention models for Holo releases on 1,024 H100 GPUs. Separately, trained on 13,824 NVIDIA GPUs in one continual-pretraining run on Summit; also worked on Frontier.

Selected Research Contributions

- **Continual LLM pretraining.** Engineered ablations using PyTorch, DeepSpeed and Megatron forks adapted to Summit's POWER9 CPUs and six-GPU nodes. Coauthored strategies matching a 10B retraining baseline with less compute (TMLR 2024).
- **Generative model efficiency.** Designed experiments, conducted ablation studies and co-wrote Würstchen (ICLR 2024 oral). Training: 24,602 A100 GPU-hours versus 200,000 GPU-hours for Stable Diffusion 2.1.
- **Pretraining data.** Designed and implemented distributed data-processing pipelines for MixtureVITae, a 422B-token corpus (TMLR 2026), and BigDocs, a multimodal dataset (ICLR 2025).

Research Experience

Research Scientist, Core Team | H Company

Sep 2024 – present

Berlin, Germany / Paris, France | Remote from Germany

- Built synthetic-data generation and desktop/web-agent evaluation pipelines and ran SFT for Holo1–3. Holo3-35B-A3B achieved 80.36% on OSWorld-Verified, ranking #1 at its April 2026 evaluation.
- Contributed to Holo and Holotron model training and collaborated with NVIDIA on official Nemotron releases.
- Contributed to Surfer-H/Holo1, Surfer 2, Holo2, Holo3, Holo3.1, Holotron and Holo4, spanning web agents and computer-use models.

Senior Research Scientist | ServiceNow Research

Jan 2024 – Sep 2024

Montréal, Canada | Trust and Governance Lab

- Built application-oriented safety evaluations for foundation models and automated red-teaming procedures for high-risk scenarios.
- Designed and implemented distributed data-processing pipelines for BigDocs, an open multimodal dataset for document and code tasks.

Postdoctoral Researcher | Mila / Université de Montréal

Jul 2022 – Sep 2024

Montréal, Canada | DAAD IFI Postdoctoral Fellow

- Researched efficient neural architectures, generative models, continual LLM pretraining and large-scale distributed training.
- Equal-contribution first author of Würstchen; coauthor of research on continual pretraining and geospatial foundation models.

Selected Publications

Simple and Scalable Strategies to Continually Pre-train Large Language Models

A. Ibrahim, B. Thérien, K. Gupta, M. L. Richter, Q. Anthony, T. Lesort, E. Belilovsky, I. Rish. TMLR, 2024.

Würstchen: An Efficient Architecture for Large-Scale Text-to-Image Diffusion Models

P. Pernias*, D. Rampas*, M. L. Richter*, C. J. Pal, M. Aubreville. ICLR, 2024. Oral presentation.

*Equal-contribution first authors. Presenting author: M. L. Richter.

MixtureVitae: Open Web-Scale Pretraining Dataset With High Quality Instruction and Reasoning Data Built from Permissive-First Text Sources

H. Nguyen et al., including M. L. Richter. TMLR, 2026. Featured Certification.

BigDocs: An Open Dataset for Training Multimodal Models on Document and Code Tasks

J. A. Rodriguez et al., including M. L. Richter. ICLR, 2025.

Surfer 2: The Next Generation of Cross-Platform Computer Use Agents

M. Andreux et al., including M. Richter. arXiv:2510.19949, 2025.

Surfer-H Meets Holo1: Cost-Efficient Web Agent Powered by Open Weights

H Company research team, including M. L. Richter. arXiv:2506.02865, 2025.

Should You Go Deeper? Optimizing Convolutional Neural Network Architectures Without Training

M. L. Richter, J. Schöning, A. Wiedenroth, U. Krumnack. IEEE ICMLA, 2021. First author.

Earlier Experience

Research Associate | University of Osnabrück

Jan 2021 – Jul 2022

- Developed methods for efficient neural architecture design and deployed face recognition and medical image models.

Data Scientist and Consultant | Freelance

Jan 2021 – Jul 2022

- Delivered computer-vision solutions for asset management and legal technology, and database migrations.

Head of Data Science | Agile. IT. Management. GmbH

Sep 2017 – Jan 2021

- Progressed from Data Scientist (2017–2018) to Head of Data Science (2018–2021). Led 10 data scientists and software engineers on projects for Adidas, Rossmann and Volkswagen Financial Services.

Artificial Intelligence Specialist | ELAN e.V.

Jul 2016 – Sep 2017

- Built a computer-vision/NLP system for copyright fraud detection in university document databases, deployed nationwide.

Education

Dr. rer. nat. | University of Osnabrück | 2019–2022. Summa cum laude.

Thesis: Towards Efficient Convolutional Neural Architecture Design.

M.Sc. Cognitive Science, 2017–2019. Distinction (1.1). University of Osnabrück.

B.Sc. Computer Science, 2013–2017; B.Sc. Cognitive Science, 2012–2016. University of Osnabrück.

Research Leadership and Service

- Secured a €92,000 DAAD IFI postdoctoral fellowship as sole applicant; awarded Google TRC and TensorFlow TPU compute grants.
- Created and taught Advanced Topics in Deep Learning, adopted into the Osnabrück master's curriculum; first supervisor of four theses and mentor to Mila graduate interns.
- Elected Mila postdoc representative; reviewed for ICLR, NeurIPS, IEEE TPAMI and other venues. Invited speaker at Microsoft Research Montréal (2023).

Languages: German (native), English (full professional proficiency).